Partial Least Squares for Essential FTIR

1st Edition.

Table Of contents

Introduction

Partial Least Squares (PLS) is a multivariate data reduction technique used in classification and quantitative analysis of infrared spectra. This manual is not meant to explain PLS itself. There are many fine sources for this information, please see the References section. This document explains how to use the PLS tool in Essential FTIR.

Installation and Licensing

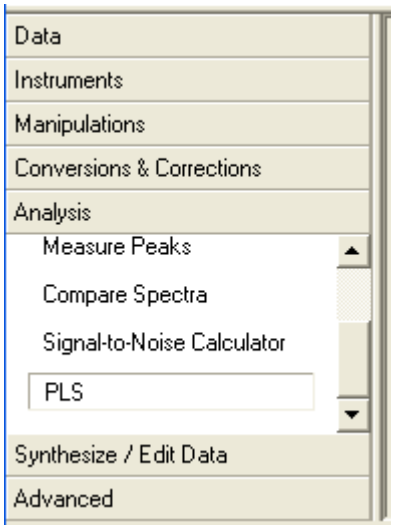
PLS is installed with Essential FTIR. The latest release of Essential FTIR can be downloaded from http://www.essentialftir.com/download.html. Installation of Essential FTIR is explained in the Essential FTIR manual, which can be downloaded from the same site.

Although included in the Essential FTIR installation, PLS is an add-on package requiring an additional license. Please contact essentialFTIR@tds.net to get a PLS license.

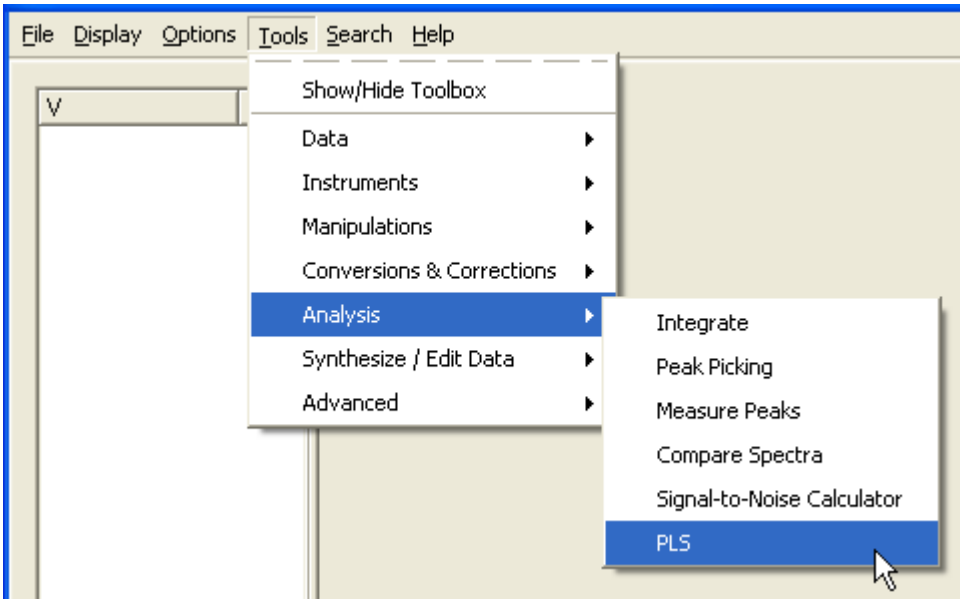
Overview

The Toolbox

The PLS tool is installed into the 'Analysis' tool category in Essential FTIR. In Essential FTIR the toolbox occupies the lower left corner of the program:



PLS can also be accessed from the 'Tools' menu:



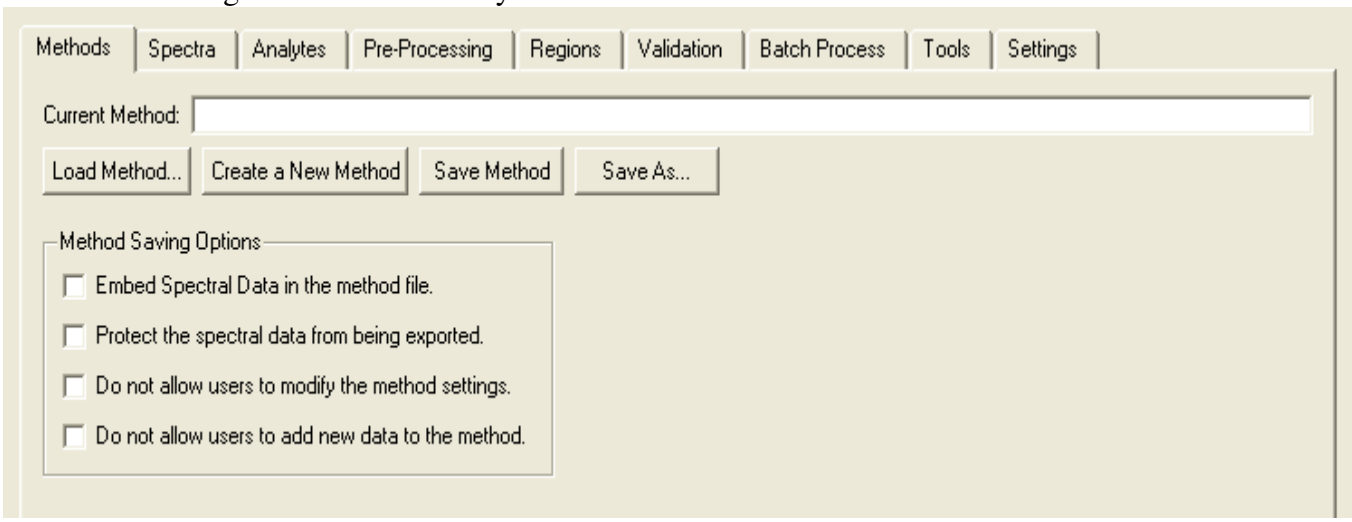
The PLS tool has these 'tabs':

Methods	Methods are created, saved and deployed
Spectra	Add calibration and test data to the method
Analytes	Define the chemical species and properties you need to quantify
Pre-Processing	Define preprocessing steps to be automatically performed during the calibration and analyses
Regions	Assign analysis regions for each analyte
Validation	Test the model against known data to find problems with the method such as spectral outliers
Batch Prediction	Using the model, analyze spectra from disk
Tools	Convenient utility functions are collected here
Settings	Various model parameters are collected here

Each of these tabs is covered separately.

The Method Tab

The method tab allows loading and saving of PLS methods. A method is the collection of data and settings that define an analysis.



Description of Buttons:

Load Method...	Load a previously saved PLS method from disk
Create a New Method	Creates a new (blank) method.
Save Method	Save the method to disk using the previously assigned filename
Save As...	Save the method to disk under a new filename.

ESSENTIAL FTIR PLS method files are saved with the extension of ".Essential FTIR_pls". At a minimum, the .Essential FTIR_pls file contains lists of all the analytes, pre-processing steps, analysis regions, and spectral data files. There are additional options for controlling what and how the method is saved. These options help in reliably and securely deploying the method and releasing it to users in the field.

The Method Savings Options:

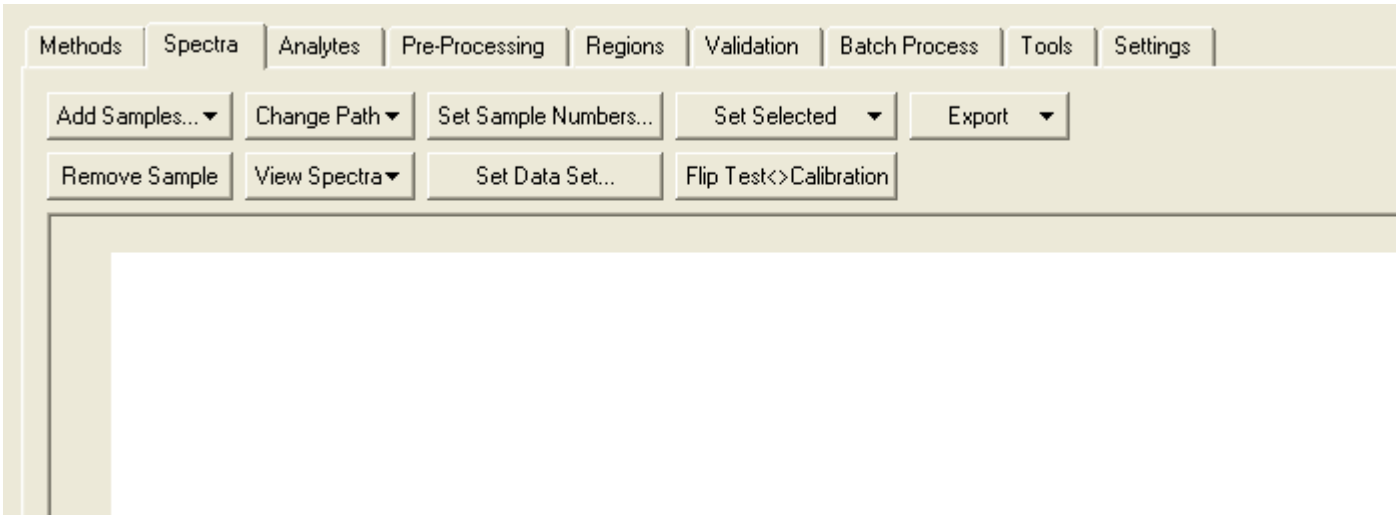
☐ Embed Spectral Data in the method file.	This puts all of the data listed in the 'Spectra' tab, that is, all the calibration and test data, directly into the .Essential FTIR_pls file. This allows you to easily transport the method and associated data to different computers
☐ Protect the spectral data from being exported.	If you choose to 'Embed Spectral Data in the method file', you can also check this box to prevent users of the method from

	exporting the spectra out of the .Essential FTIR_pls file to individual spectral data files. This allows users in the field to modify the method, and examine the spectral data, but the individual calibration and test spectra cannot be exported from the .Essential FTIR_pls file. This protects your investment in the original data that comprises the method. Also, the data placed in the .Essential FTIR_pls file is encrypted. Be careful with this option because; always maintain a backup of your original datafiles because this step is irreversible and even the makers of Essential FTIR cannot recover protected data.
☐ Do not allow users to modify the method settings.	When deploying a method, you may not want to allow users to change anything associated with the method.
☐ Do not allow users to add new data to the method.	As new data is acquired, that data can be added to the method. If you check this box, users in the field cannot add new data to the method.

These deployment options are irreversible. Therefore, if you are the method developer, you should not use any of these options while developing the method. If you lock down the method, you will not be able to make any changes to it.

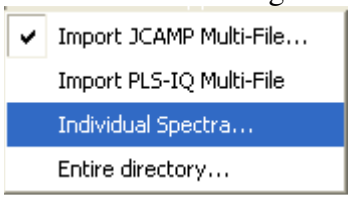
The Spectra Tab

Adding spectra and their associated concentration information is by far the most tedious, time-consuming and error-prone part of creating PLS methods. Often the required information is in different files, and those files may be in different formats. For instance, the spectral files of calibration standards may be in any number of formats, and the concentration information about those spectra is usually stored externally in another file, often an Excel spreadsheet or tablulated text file. Chemometric practitioners often use a combination of Excel, scripting languages, batch files, and various specialized editors to manage this information and insert it into analysis programs. Essential FTIR tries to make this process as easy as possible by including a number of features to simplify the tasks that are involved.



Description of Buttons:

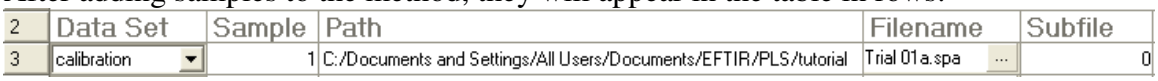
The downward facing arrow on this button gives a menu of choices:



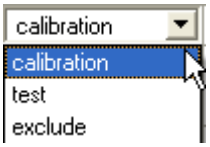
Import JCAMP Multi-File	Collections of data are sometimes distributed in what are called multi-files, which are many spectra (sub-files) contained in one large file, hence the name 'multi-file'. JCAMP multi-files can also

	contain information about the concentrations of analytes for each spectrum, so this is an easy way to enter all that information from a single file.
Import PLS-IQ Multifile	PLS-IQ is the name of a PLS program from Galactic Industries, now part of Thermo Scientific. This option allows you to easily import existing data from PLS-IQ
Individual Spectra	Select one or more files using the Essential FTIR multi-selection file browser
Entire Directory	Choose a directory and import all of the spectra in that directory.

After adding samples to the method, they will appear in the table in rows:



Under the 'Data Set' column is a drop-down list that includes these items:



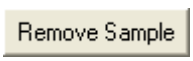
The spectrum can be assigned to the calibration set, the test set (this is covered in the 'Validation' section), or excluded from both test and calibration sets.

The 'Sample' column contains the sample number. Often, replicate measurements of the same sample, or of samples of the same concentration, are included in the method to model instrument-to-instrument variation or sample mixing variation. Mathematically, it is necessary to treat these replicate samples as a group. This column displays, and can be used to edit, the sample number assigned to a spectrum.

The 'Path' column tells what directory the spectral file is in. This is needed because while developing the method you may switch between poplulations of data that share the same filenames, but are in different directories on the computer. (The 'Change Path' button allows you to switch between directories).

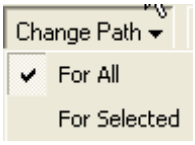
The 'Filename' column contains the root name of the spectral datafile, without the path.

The 'Subfile' column contains the subfile number when the spectrum is from a multifile. In this case, the spectrum is in a single-spectrum datafile, and the subfile is 0.

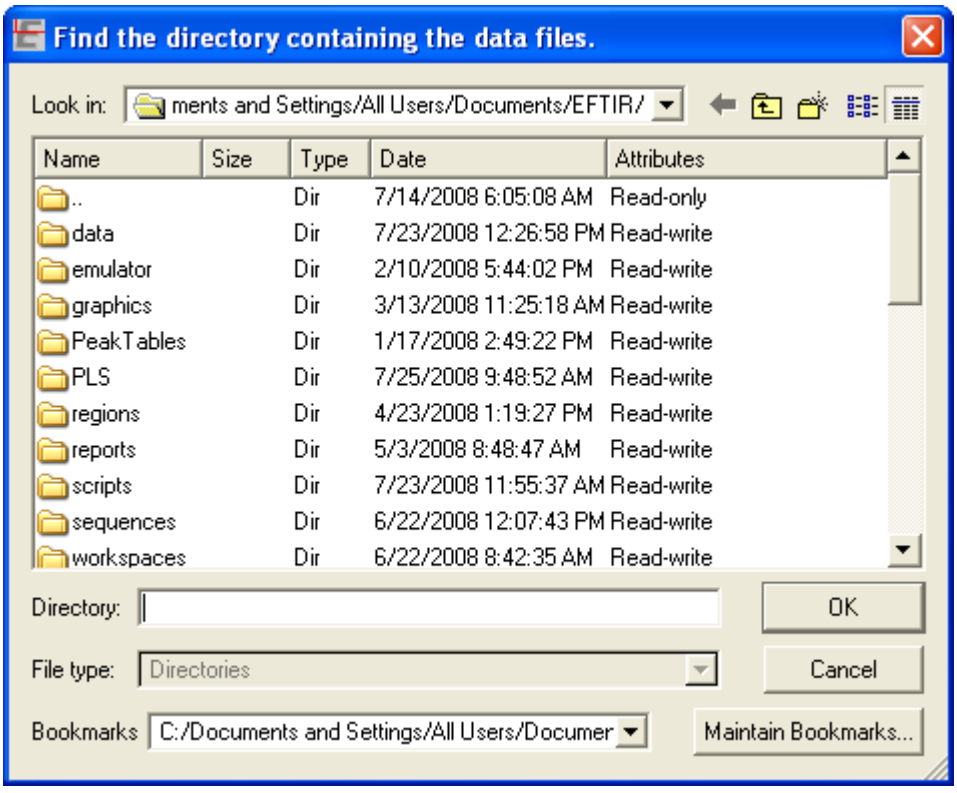


After selecting a row or multiple rows in the table of spectra, click this button to remove the spectrum from the method. Select single rows in the table by left-clicking on the left-most column of numbers in the table. To select multiple rows, left-click and drag the mouse in the left-most column of numbers.

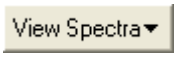
The arrow signifies there is a menu of choices for this operation:



While developing and testing a method, you may switch between populations of data that share the same filenames, but are in different directories on the computer. This Change Path function allows you to modify the path to the selected spectra. The usual Window 'Directory Selection' dialog will appear:

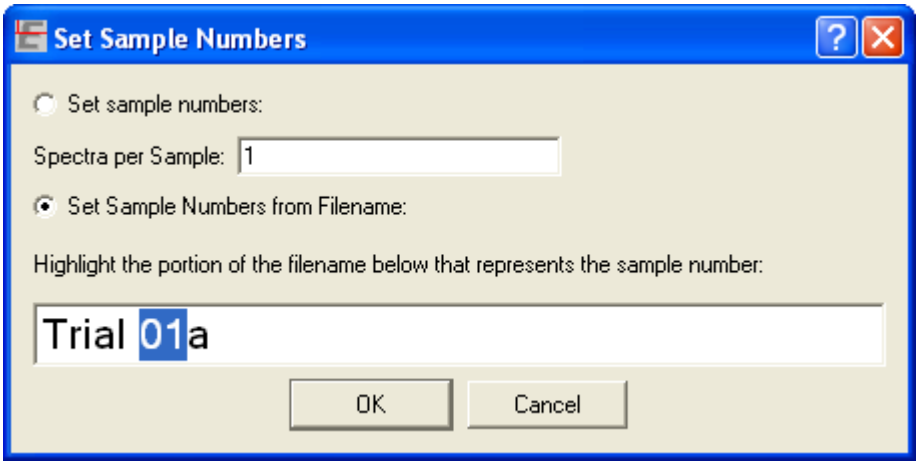


Take care to choose a directory that actually contains the files you are changing the path to, or you will get read errors when trying to use those files.



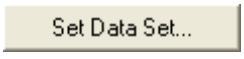
This button allows you to view the selected spectra, or all spectra, in an Essential FTIR workspace. See the Essential FTIR manual for a complete description of the data workspaces and how change the display of spectral data.

Clicking this button will cause this dialog to appear:

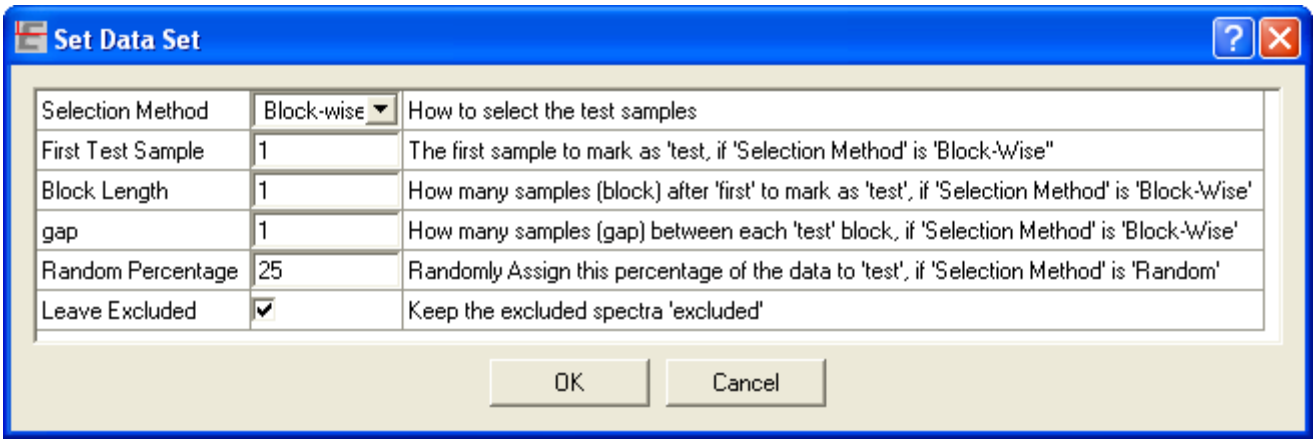


This dialog is used to assign the numbers in the 'Sample Number' column in the spectra table. On this dialog are two options for setting the sample numbers associated with spectra in the spectra table. 'Set sample numbers' radio button merely numbers the spectra sequentially, starting at 1, and grouping the spectra into sample groups using the 'Spectra per Sample' setting. For instance, if this is set to '3', the first three spectra in the table will be given sample number 1, the second three spectra sample number 2, and so on.

Very often, the file names contain the sample number information, but that information can be anywhere in the filename, in different formats. The second choice on this dialog, 'Set Sample Numbers from Filename', allows you to select, in a sample filename, the portion of the name that includes the sample number. In this case, it is the '01' portion of the filename. If you select this choice the portion of the filename you select as representing the sample number is used to create a pattern that is then used to extract the sample number from the filename for every spectrum in the spectra table.



Clicking on this button causes this dialog to appear:



This is used to assign the values in the 'Data Set' column for spectra in the table, that is, to assign spectra to be in the 'calibration' or 'test' data sets.

Keep in mind that spectra are grouped into 'samples', and spectra with the same sample number are treated by the software as an indivisible group. If you had two spectra with the sample number '1', you cannot assign one of them to the test set and the other to the calibration set. The numbers in this table for 'First Test Sample', 'Block Length', and 'gap' refer to sample numbers, not individual spectra.

Selection Method: this drop-list has the values: Block-wise or Randomly. If you choose 'Randomly', the 'Random Percentage' determines how many samples are assigned to the test set.

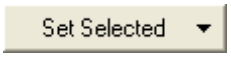
Usually the data sets are assigned using 'Block-wise' selection, where again a 'block' is a group of spectra with the same sample number. Here are some examples:

To assign every other 'sample' to 'Test', starting with sample 1:
First Test Sample: 1; Block Length: 1; Gap: 1.

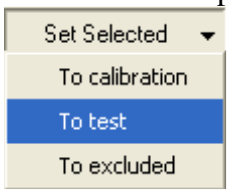
To assign all samples to 'Calibration':
First Test Sample: 1; Block Length: 0; Gap: 1.

To assign all samples to 'Test':
First Test Sample: 1; Bock Length: 1; Gap: 0.

To assign test / calibration samples in the ratio of 1:2, starting with sample number 2:
First Test Sample:2; Block Length: 1; Gap: 2



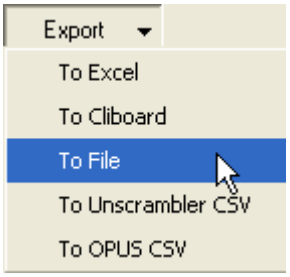
The buttons' drop list has these selections:



The selected row or rows in the spectra table can be manually assigned to a data set using this function.

Spectra assigned to 'test' become the calibration set, and those assigned to calibration become test. Any excluded spectra are left alone.

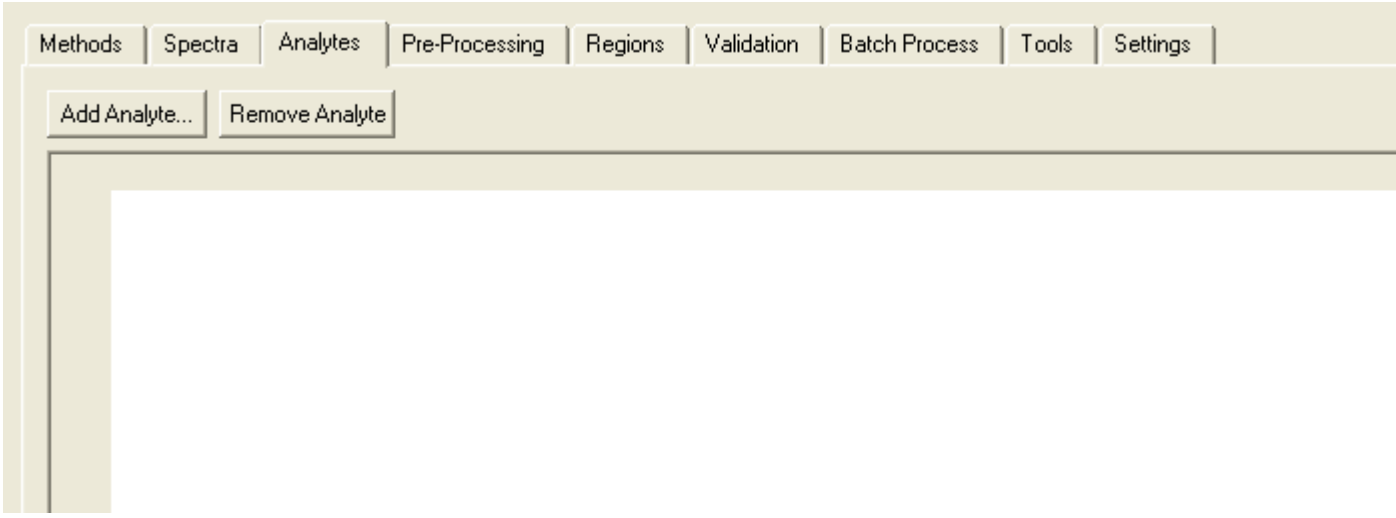
The menu choices for this button are:



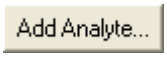
This exports all of the data in the table directly into Excel; to the Windows Clipboard, or to a .csv (comma-separated-value) file. The last two choices allow you to interchange data with Unscrambler or OPUS. Unscrambler is a multivariate statistics package, which includes PLS and is very popular among chemometricians. OPUS software is from the FTIR manufacturer Bruker. For these two options, the spectra table is exported to the windows clipboard with the correct column information to allow direct pasting (via the clipboard) into these other software packages.

The Analytes Tab

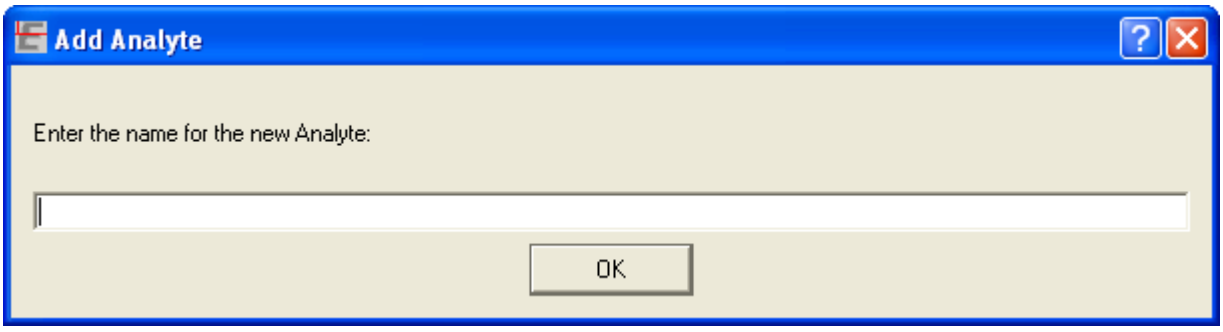
An analyte is a chemical compound to include in the analysis method.



On this tab you manage the Analytes, which are the chemical species, sometimes referred to in this document as 'compounds'. These are the species that the PLS method is being used to analyze the composition of unknown spectra for.



This dialog appears:



You simply give the analyte a name.

After adding an analyte, in this case 'Water', the table will look like this:

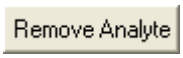
	Analyte:	Status:	Factors:	Units:
2				
3	Water	include	10	?

The 'Analyte' column has the name you gave on the dialog. You can edit it directly in the table by double clicking in the cell and editing the name.

Status: This droplist has 'include' and 'exclude' as the choices. You can exclude an analyte from the calibration and prediction by setting this to 'exclude'.

Factors: this is the number of factors to use in calibrating this analyte. The default value is 10, which is probably too high for most situations. This number is 'tuned' in the Validation step.

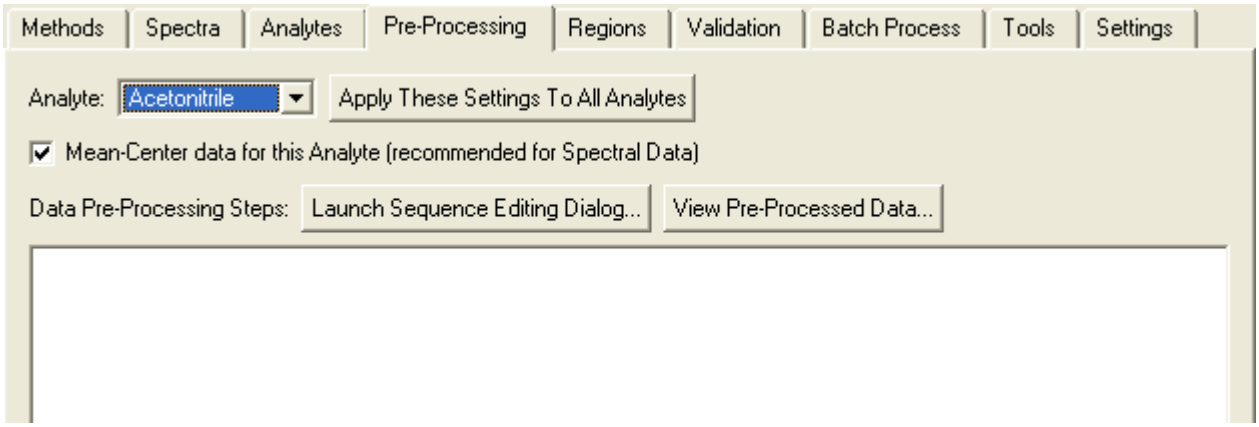
Units: the units to label concentration values with for this analyte. You can edit directly by double clicking in the cell.



The selected row or rows in the analyte table will be removed from the method. Before they are removed, you will be asked if you really want to do this.

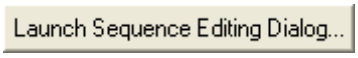
The Pre-Processing Tab

Data can be processed in various ways before the quantitative analysis is performed, in order to bring out the information in the data.

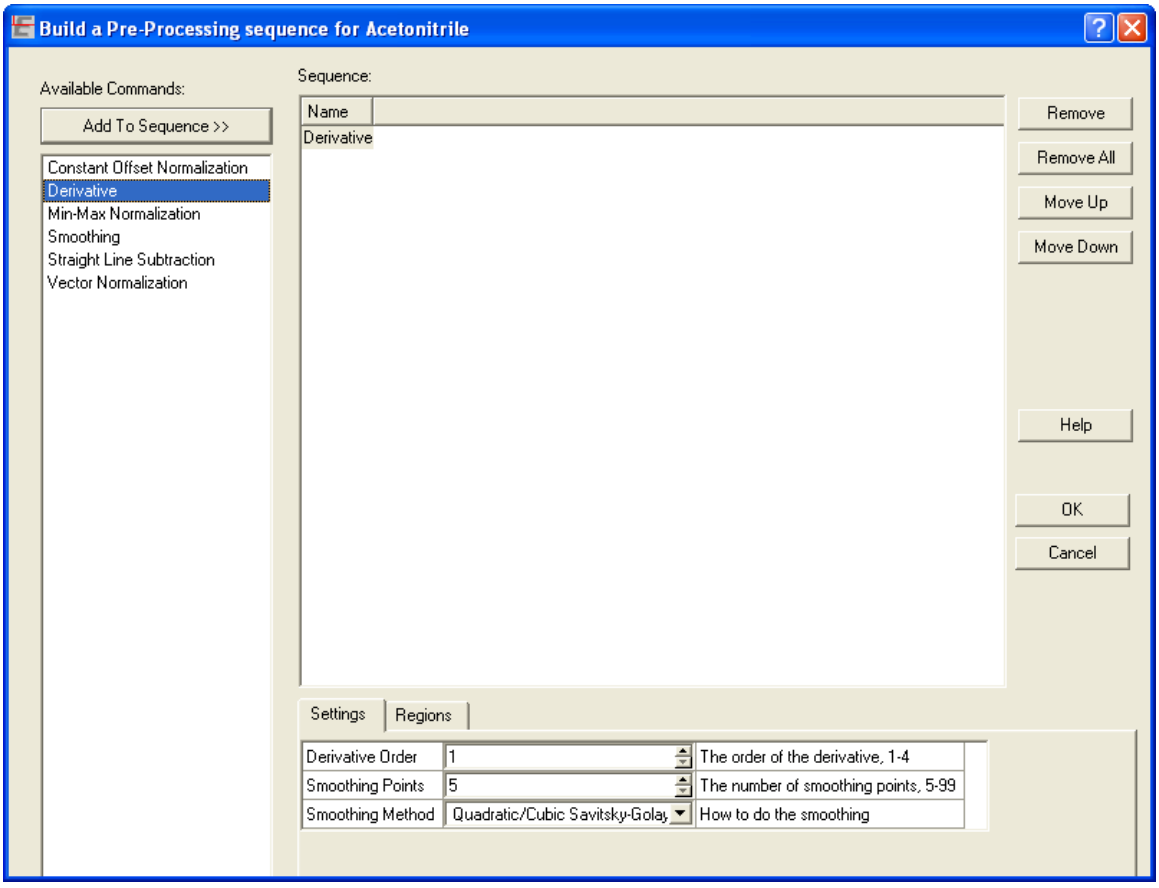


Each analyte can have separate pre-processing steps defined.
Choose the analyte you want to work with from the 'Analyte' list.

The Mean-Center check box is checked by default because this operation is appropriate for most cases using spectral data. If you do not want to mean-center the data, uncheck this.



This dialog will appear:

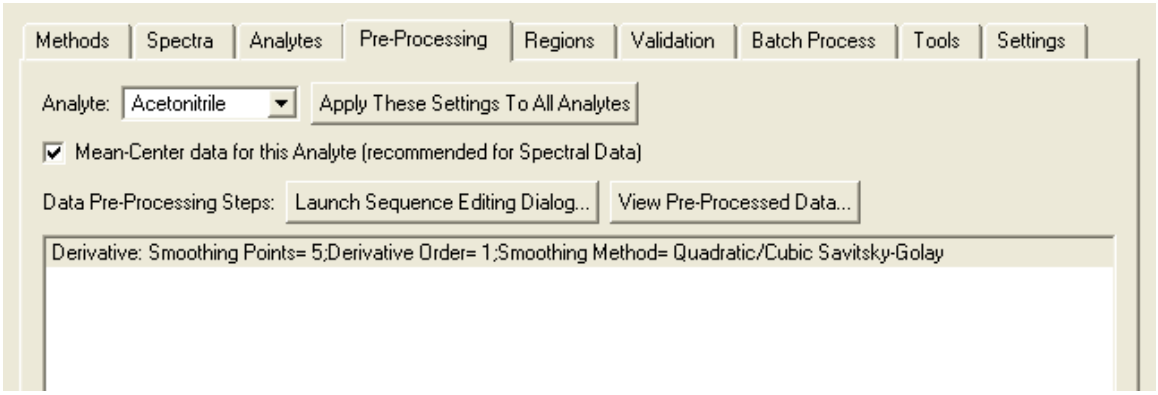


This dialog is very similar to the 'Batch Sequence Editor' in Essential FTIR, please see the Essential FTIR manual for more information. This dialog just presents a certain subset processing operations that are useful for PLS pre-processing.

On the left side of the dialog is a list of the available pre-processing operations. The list in the middle of the dialog is the operations that have been added to the pre-processing sequence for the given analyte.

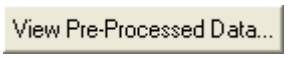
To create a pre-processing sequence, highlight an Available Command (in the figure above, 'Derivative' has been highlighted. Then click 'Add To Sequence' to put it into the 'Sequence' list. The settings and regions tab allow you to set parameters needed for any particular step highlighted in the Sequence list. When done, click 'OK'. The sequence will be automatically added to the PLS method for the analyte.

In this example, after clicking OK, the 'Pre-Processing' tab for the analyte will look like this:

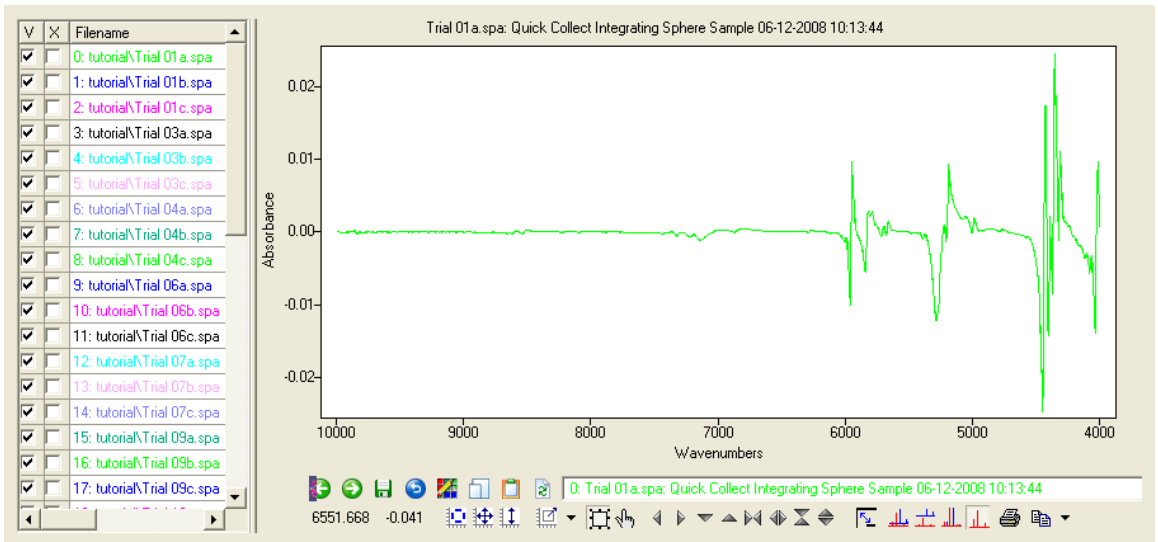


A summary of the pre-processing steps is displayed.

Usually you want to apply the same pre-processing steps to all the analytes. The 'Apply These Settings to All Analytes' button will propagate the settings for the selected analyte to all the other analytes in the method.

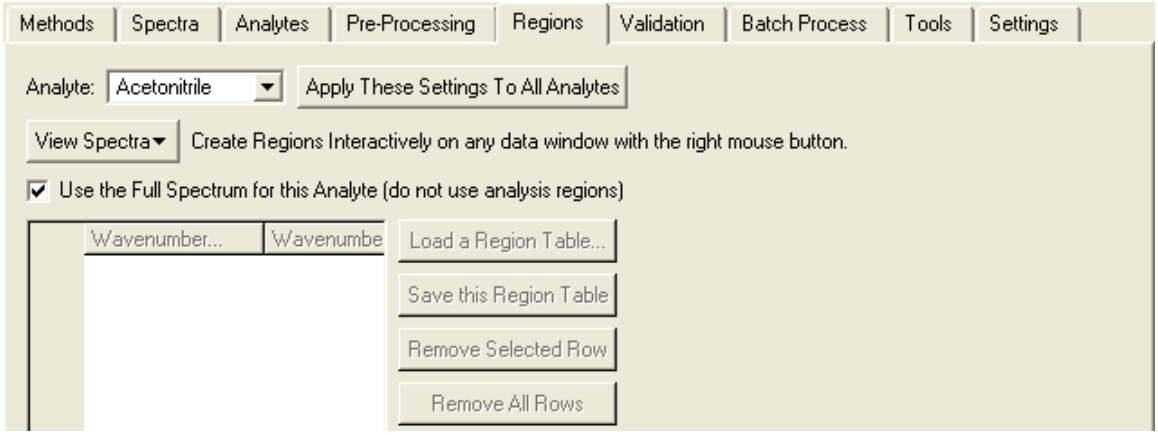


To see what the pre-processing will do to the data, click this button. All of the spectra in the Spectra table will be pre-processed and displayed in a pop-up dialog that embeds an Essential FTIR workspace:



The Regions Tab

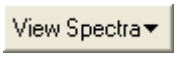
Analytes usually have spectral regions that contain information specific to that analyte. Specifying spectral regions, rather than full-spectrum analysis, can improve the results and make the analysis faster.



On the Regions tab, you assign analysis regions for each analyte.

Regions are assigned interactively in a spectral display workspace; please see the Essential FTIR manual which has an entire section about creating and changing region selections. Basically, you create regions markers by right-clicking in the spectral display window. For PLS you can create as many regions as you need. Region markers can be moved by dragging with the left mouse button and removed by clicking on it with the right mouse button.

The creation of regions is covered in detail in the tutorial section of this document.



To assign analysis regions for an analyte, you need to display some spectra. This button places data from the spectra table into a workspace.

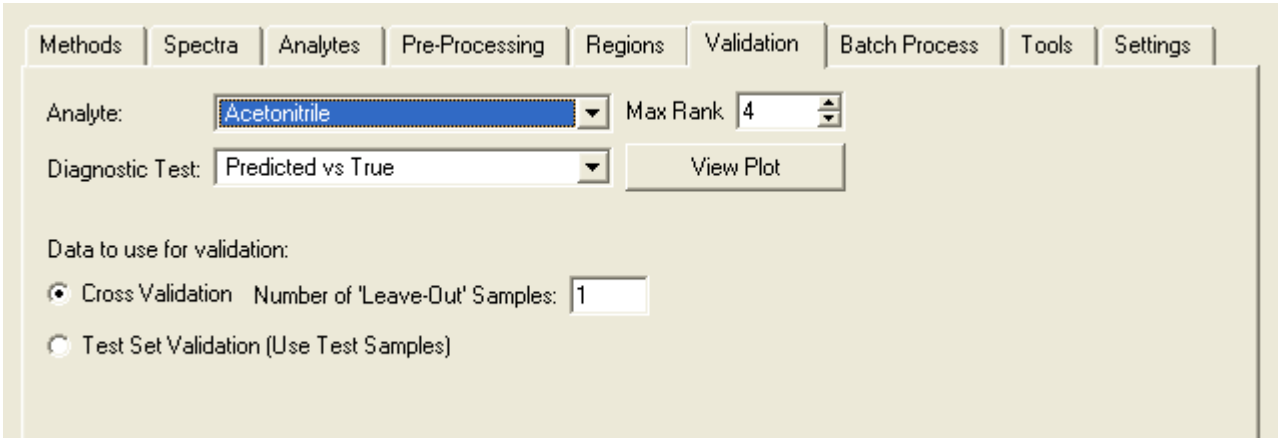
You may want to apply the same analysis regions to all analytes in the method. This may seem an unlikely thing to want to do, but there may be regions of the spectra that you want to exclude for all the analytes, for instance the spectra may contain information from beyond the detector cut-off.

This check box allows you to toggle between full-spectrum and region analysis. Even if you have defined analysis regions, you can tell the software not to use them.

These buttons are covered in detail in the Essential FTIR manual, but their functions are self-explanatory. The same functions are available by clicking on the 'Wavenumber…' column headers in the wavenumber table.

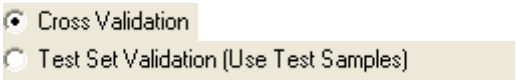
The Validation Tab

Validation means testing the PLS method against known data.

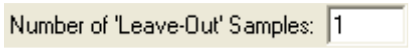


The functions on this tab are used to tune the method for optimal performance. The purpose of validation is to find the optimum number of factors for an analyte and to find outlying samples in the method. Samples can be concentration outliers or spectral outliers.

First you must select which data to use for the validation.

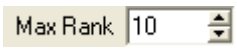


Validations are performed using all of the calibration spectra in a 'Cross Validation', or using the spectra identified as 'Test' spectra on the spectra table. Cross validations can be very time consuming for large datasets. A number of samples in the calibration set are removed from the calibration set, which number is set in this edit field:



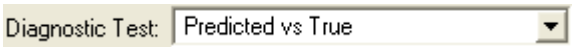
The method is calibrated without those N samples, and the N removed samples are treated as unknowns and their concentrations are predicted using that calibration. Then those temporarily removed calibration samples are put back into the calibration set, and the process is repeated for every sample.

Test set validation, on the other hand, only has to perform one calibration, using the calibration data set, and then all of the test set spectra are predicted against that calibration.



For any given diagnostic test, results are computed up to 'Max Rank'.
The results for all ranks up to 'Max' can be displayed without forcing a re-calibration.

After selecting the test to perform in the 'Diagnostic Test' list,



Click the 'View Plot' button

to perform the selected test and to display the test results.

Equations for the diagnostics:

In the following equations, M is the number of samples in the training or validation set, f is the number of factors, Cp is the predicted concentration, Ck is the known concentration. $\overline{Ck}$ is the average of the known concentrations.

SSE

Sum of Square Errors: The sum of the squared concentration residuals.

$$SSE = \sum_{i=1}^{M} (Cp_i - \overline{Ck})^2$$

R-Squared: R^2

The Coefficient of Multiple Determination : sometimes called Explained Variance. There is variance in the known concentration values; this tells how much of that variance is reproduced by the predicted values. Higher values indicate a better correlation.

$$R^2 = 100 \cdot \left[1 - \frac{SSE}{\sum_{i=1}^{M} (Ck_i - \overline{Ck})^2} \right]$$

RMSECV: Root Mean Square Error of Cross Validation
RMSEP : Root Mean Square Error of Prediction.

This measures the precision of the test analysis.
These are calculated using the same formula, except that in RMSECV the Cp are from cross validation and in RMSEP the Cp are from calculated from the test samples.

$$RMSECV, RMSEP = \sqrt{\frac{\sum_{i=1}^{M} (Ck_i - Cp_i)^2}{M}}$$

SpecRes: SpecRes (sometimes called RMS(residual)) is the square root of the sum of the square of the spectral residual, where the spectral residual is the difference between the original spectrum and the theoretical spectrum derived from the PLS factors. The residual can be created additively, by adding in the contributions of each factor, or by subtracting the contribution of each factor from an unknown spectrum. In the PLS prediction, the contribution of each factor is subtracted from the spectrum, and the spectrum that is left at each step of this process of iterating over the factors is called the 'spectral residual'. This is a measure of how well the method is modeling the spectral data, and can be used to spot spectral outliers. Samples with spectral residuals greater than the other samples may be spectral outliers and should be excluded from the method.

Smaller values are better, and indicate that the model explains more of the spectral structure. This value can be used to identify spectral outliers. In the following formula, M is the number of frequency values in the spectrum, and the residual is calculated over all the frequencies, denoted by subscript 'j'.

$$Specres = \sqrt{\sum_{j=1}^{M} (Measured_j - \mathrm{Re}constructed_j)^2}$$

Mahalanobis Distance

This measures the scaled distance of a sample point from the mean of all the remaining points. The distance is scaled in all dimensions by the range of variation in the points. The Mahalanobis Distance measures the reliability of an analysis.

In the following equation, X is the matrix of spectral scores from the calibration; there is one for each calibration spectrum at each factor. t are the scores calculated during the prediction step for an unknown. Superscript T stands for matrix transposition; superscript -1 means matrix inversion. Subscript 'i' is the factor number.

$$Mah.Dist._i = t_i^T (X^T \cdot X)^{-1} \cdot t_i$$

The Mahalanobis Distance tells how well a spectrum matches the spectra in the calibration set. It is in units of standard deviations from the centroid comprised of all the calibration spectra.

The available diagnostic tests

The interpretation of these will be discussed in the tutorial section.

Predicted vs. True: This is the first test that is usually done because of its simplicity and ease of interpretation. The predicted concentration values are plotted against the actual concentration values for each spectrum. This can be used to spot concentration outliers.

Difference vs. True: The difference between the predicted and actual concentrations is plotted for each spectrum (that is, the concentration residual). This is useful in spotting concentration outliers; samples with larger concentration residuals are probably concentration outliers and should be excluded from the calibration.

Sample Number vs. SpecRes: See above for a description of SpecRes. This test is used to spot spectral outliers. Smaller values of SpecRes are better.

Sample Number vs. Mah. Dist..: 'Mah. Dist' is Mahalanobis Distance and is described above. In Essential FTIR, any point with a Mahalanobis Distance more than three standard deviations from the mean of the others is flagged as an outlier.

Sample Number vs. Score:
The 'Score' is a measure of how much

Rank vs. R2:
Rank is the factor number, and R2 is R-Squared, described above.

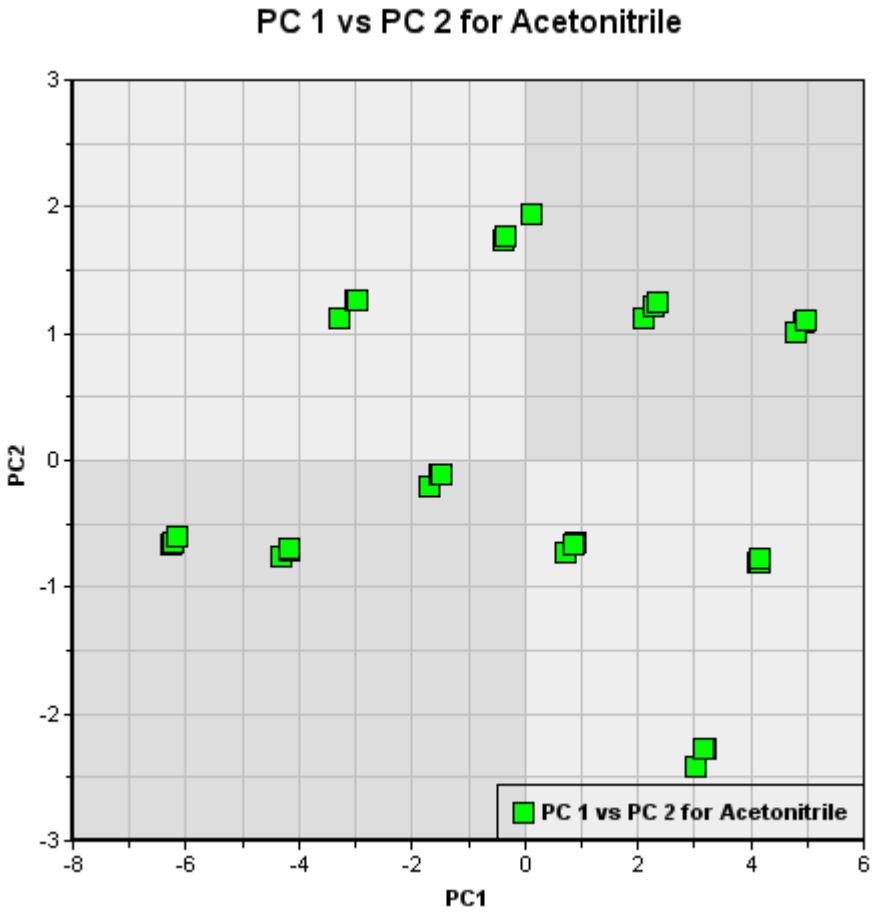
Rank vs. RMSE[CV,P]:
RMSECV is 'Root Mean Square of Error for Cross Validation' and RMSEP is 'Root Mean Square of Error in Prediction'. RMSEP is obtained when doing a test set validation' RMSECV for Cross Validation.

RMS(residual) vs. Mah. Dist.:
The spectral residual is plotted against the Mahalanobis distance. This is useful for spotting outliers.

Calibration Scores:
Prediction Scores:

These are 'score plots', using the PLS scores for the Calibration or Prediction spectra. This diagnostic allows plotting one factor's score against another factor. The score plot is examined to see if is consistent with the data, because it shows the relationship of samples to each other in the new variable space. For instance, in this plot of Factor 1 vs.

Factor 2 for the tutorial example, which contains 3 replicates of each sample, notice how the replicates group together. It would indicate a problem if they did not. Also, the samples with concentrations of in the middle of the concentration range are in the center of the plot.

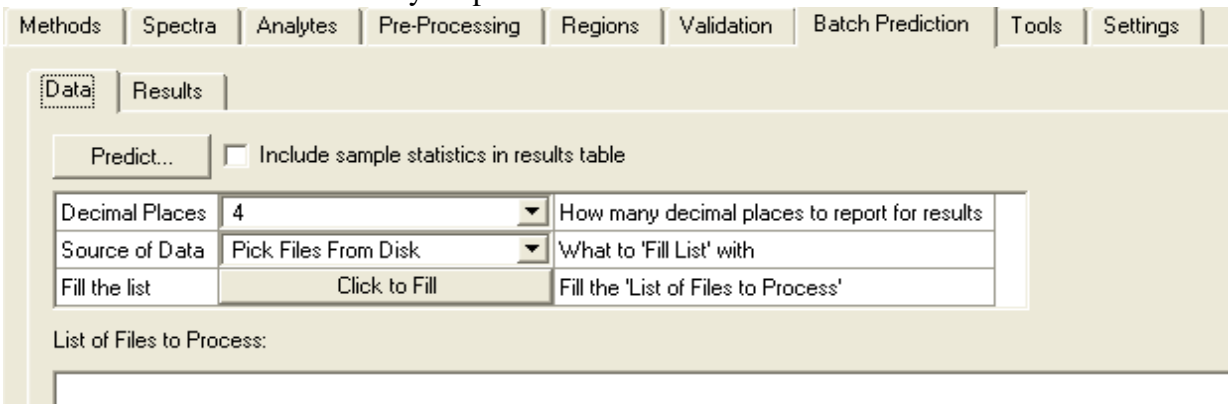


Weight Vector:
Loading Vector:
Regression Coefficients:

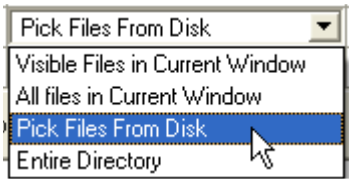
These take the form of spectra because the X axis of these is the same as the calibration spectra. These are the actual calibration data from the PLS algorithm, and they allow you to examine how PLS 'sees' the data; that is, how it is modeling the spectra.
Typically, at higher factor numbers, these 'spectra' look more and more noisy. As the number of factors in the analysis increases, more of the noise in the system is being modeled.

The Batch Prediction Tab

Batch Prediction is used to analyze spectra from disk files.



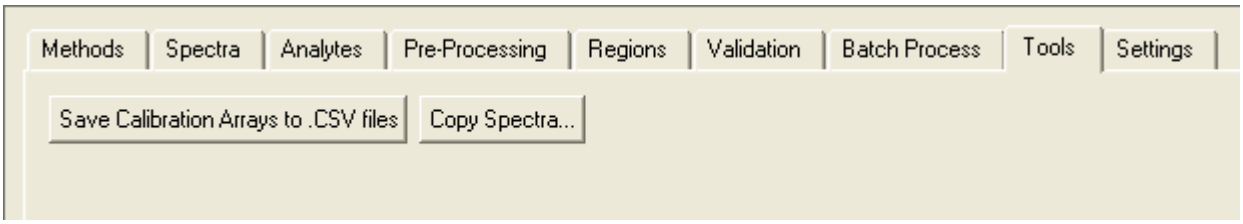
This tab allows you to analyze disk files using the method.
The 'Source of Data' allows you to choose which files to analyze:



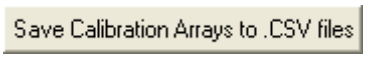
The 'Include sample statistics in results table' will include Mahalanobis distance, Spectral Residual, F-Value and F-Prob statistics for each sample.

The Tools Tab

Miscellaneous utility functions are collected here.



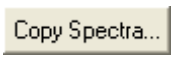
Description of Buttons



All of the internal calibration matrices that are computed are saved as Comma-Separated Value (CSV) files and placed in the same directory as the method file. After the operation, a dialog summarizing the files that were created is displayed. This feature is very useful if you want to import the arrays into a different program, or to use the calibration matrices to perform predictions in a different program. Together, these arrays are everything that is needed to perform validation and prediction of samples.

Using Acetonitrile from the tutorial as an example, the following files are created. The PLS terminology is always confusing, different texts use different terms.

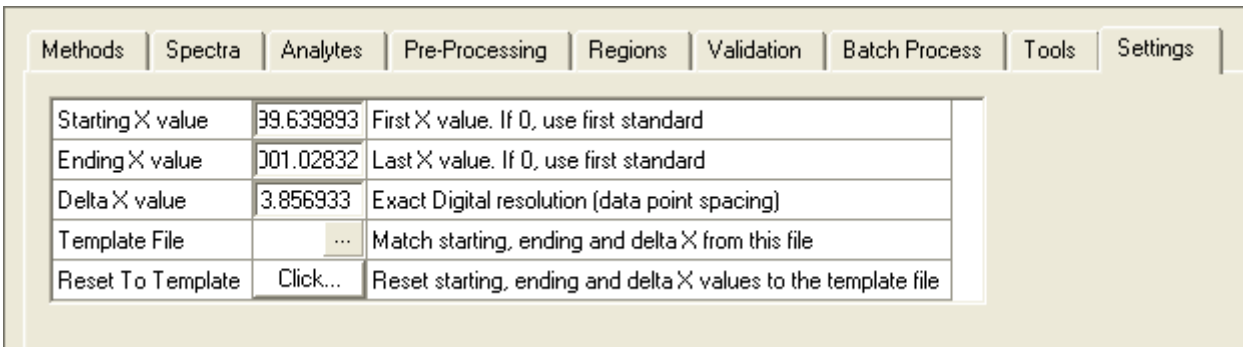
Acetonitrile_W.csv	The weight loading vector.
Acetonitrile_T.csv	The spectral scores vector (aka 'latent variable')
Acetonitrile_V.csv	The chemical loadings.
Acetonitrile_B.csv	The spectral loadings.
Acetonitrile_Bhat.csv	For 'short prediction', see Martens p. 122
Acetonitrile_b0.csv	For 'short prediction', see Martens p. 122
Acetonitrile_cmean.csv	The mean-centered concentrations.
Acetonitrile_xmean.csv	The mean-centered spectra.



Use this to copy all of the files referenced in the 'Spectra' tab to another directory. This is a way to aggregate all of the data to a single location. Also, sometimes users will want to 'branch' the spectra and modify them in ways not available in the PLS tool. You can copy the spectra, modify them, and then use the 'Change Path' button on the Spectra tab to use the modified data in the method.

The Settings Tab

Sometimes the spectra used in a PLS method are from different sources, and have to be made compatible. This table of settings controls the spectral range and digital resolution of the data. By default, when the first spectrum is added to the method, these values are set to reflect that spectrum.



Tutorial

The tutorial data is available as a separate download. Please download setup_eftir_pls_tutorial.exe from http:\\www.essentialFTIR.com/tools.html and install it on your computer.

The data is installed into the directory "C:\Documents and Settings\All Users\Documents\ESSENTIAL FTIR\PLS\tutorial". The actual location of this may be different depending on your operating system; the so-called 'Shared Documents' folder is called different things in different versions of Windows.

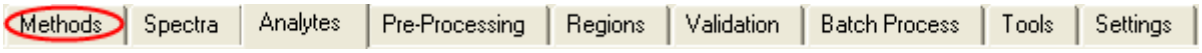
PLS method files are given the extension ".eftir_pls". In the tutorial directory is a file named 'tutorial.eftir_pls' which you can load directly into Essential FTIR. However, the purpose of this tutorial is to teach you how to create a method from the beginning.

The tutorial method is made up of Near Infrared spectra of a mixture of Water, Methanol and Acetonitrile. There are 21 mixtures with different concentrations of these, and three repeat measurements were made of each mixture.

Create a new method.

You could just load the method file that is installed with the tutorial, named tutorial.eftir_pls, but the goal of this tutorial is to show how to create a method from the beginning.

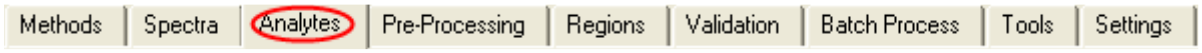
Click on the Method tab.



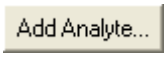
Click on the 'Create a New Method' button.

Add analytes.

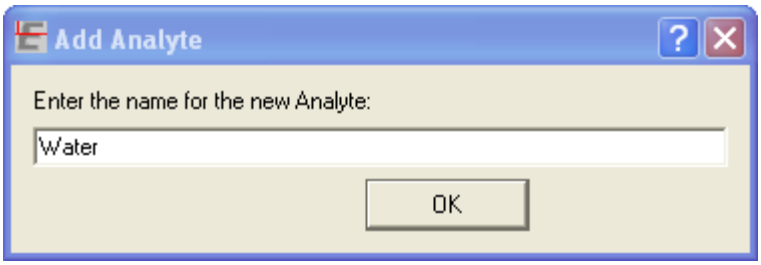
Click on the Analytes Tab.



Add Water as an analyte.
Click on the 'Add Analyte…' button:



In the dialog that appears, type 'Water', and click 'OK'.



Do the same steps to add 'Methanol' and 'Acetonitrile'. It is important to add them in this order, because later we will paste in a concentration table that is formatted this way. After adding Water, Methanol and Acetonitrile, the Analytes table will look like this:

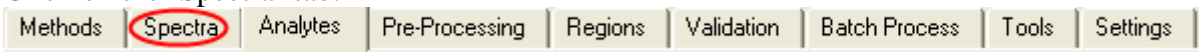
	1	2	3	4	5
1	C:/Documents and Settings/All Users/Documents/EFTIR/PLS/tutorial/tutorial 1.eftir_pl				
2	Analyte:	Status:	Factors:	Units:	Mah. Limit
3	Water	include	10	?	3.0
4	Methanol	include	10	?	3.0
5	Acetonitrile	include	10	?	3.0

In this table, you can directly edit the properties associated with the analyte. In particular, you will have to establish the correct number of Factors (aka 'Rank') to use for each analyte. This is discussed in the tutorial section. The 'Mah. Limit' column specifies the Mahalanobis Distance limit for the analyte. For all analyzed samples, a Mahalanobis Distance is computed, and if that distance is greater than the limit, the sample is flagged as an outlier, meaning that the analysis may not be valid for that sample.

Adding Spectra.

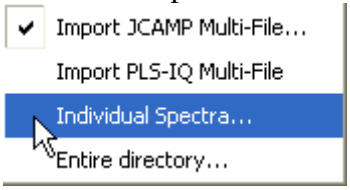
This is without a doubt the most tedious and error-prone step. Fortunately, Essential FTIR has feature to minimize the tedium.

Click on the 'Spectra' tab.

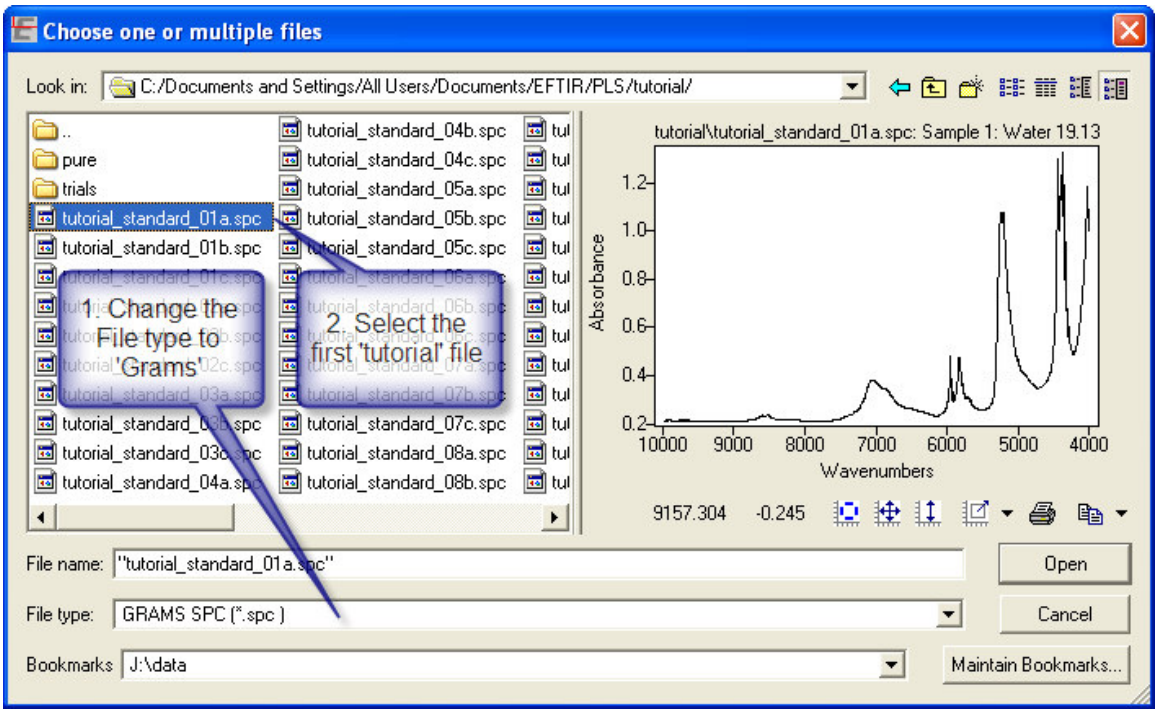


Click on the 'Add Samples' button.

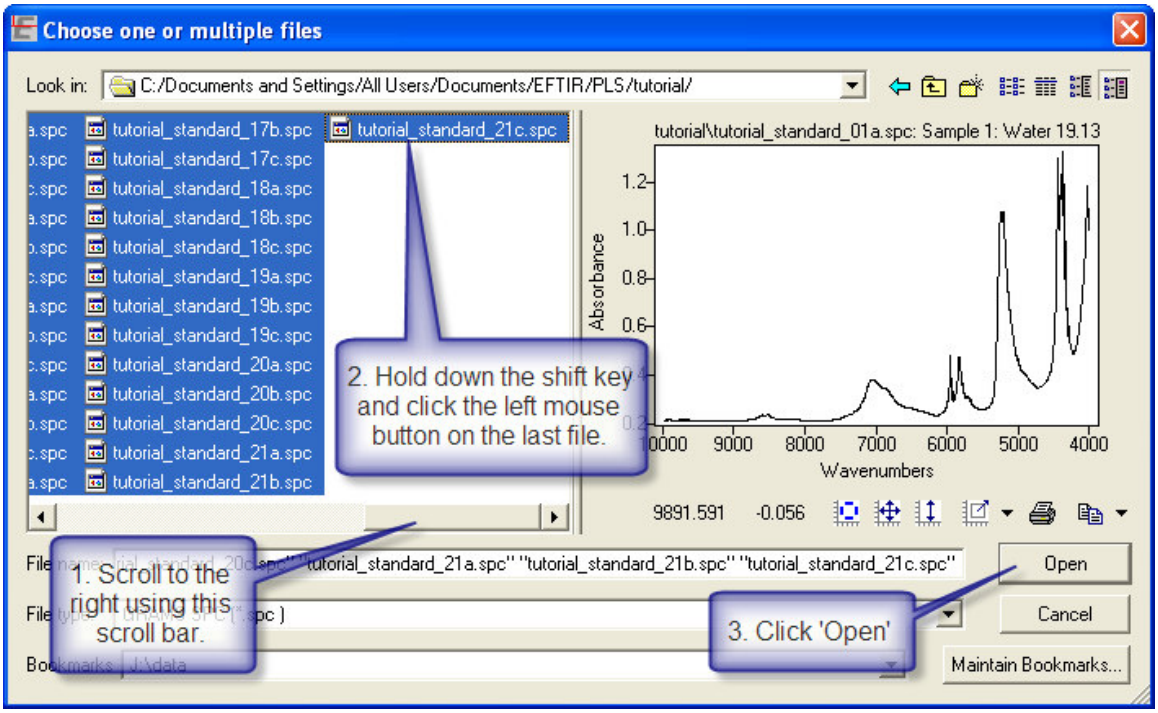
From the drop-down menu, click 'Individual Files':



The Essential FTIR file selection dialog will appear. To simplify things, set the 'File type' to 'Grams', and then select the first tutorial file, 'tutorial_standard_01a.spc'.



Next, use the horizontal scroll bar on the file list to scroll to the end, hold down the Shift key and click the left mouse button on ‘tutorial_standard_21c.spc’. Then click the ‘OK’ button:



Again, it is important to get the files in the right order. The spectra table will now contain 63 spectra and look like this:

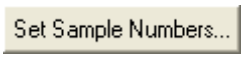
	1	2	3	4	5	6	7	8
1								
2	Data Set	Sample	Path	Filename	Subfile	Water	Methanol	Acetonitrile
3	calibration	1	C:/Documents and S	tutorial_standard_01a.spc	0	m	m	m
4	calibration	1	C:/Documents and S	tutorial_standard_01b.spc	0	m	m	m
5	calibration	1	C:/Documents and S	tutorial_standard_01c.spc	0	m	m	m
6	calibration	1	C:/Documents and S	tutorial_standard_02a.spc	0	m	m	m
7	calibration	1	C:/Documents and S	tutorial_standard_02b.spc	0	m	m	m
8	calibration	1	C:/Documents and S	tutorial_standard_02c.spc	0	m	m	m
9	calibration	1	C:/Documents and S	tutorial_standard_03a.spc	0	m	m	m
10	calibration	1	C:/Documents and S	tutorial_standard_03b.spc	0	m	m	m
11	calibration	1	C:/Documents and S	tutorial_standard_03c.spc	0	m	m	m
12	calibration	1	C:/Documents and S	tutorial_standard_04a.spc	0	m	m	m
13	calibration	1	C:/Documents and S	tutorial_standard_04b.spc	0	m	m	m
14	calibration	1	C:/Documents and S	tutorial_standard_04c.spc	0	m	m	m
15	calibration	1	C:/Documents and S	tutorial_standard_05a.spc	0	m	m	m
16	calibration	1	C:/Documents and S	tutorial_standard_05b.spc	0	m	m	m
17	calibration	1	C:/Documents and S	tutorial_standard_05c.spc	0	m	m	m

Part of the table is highlighted in yellow. These are the concentration information. The table is filled with 'm' because those values are missing. Any cell in the table with an 'm' (any non-numeric character(s) will do, such as '?' or 'N/A' or 'missing') is marked as 'missing' and that spectrum will be excluded from the data set for that Analyte.

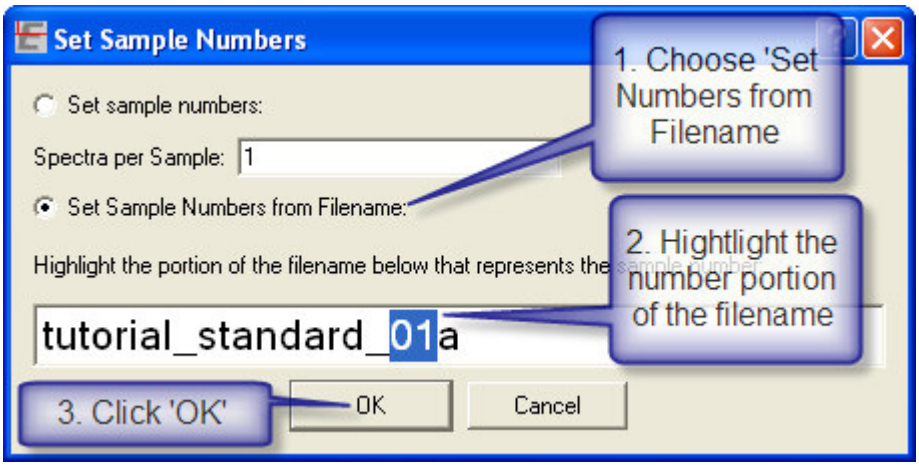
Entering the Sample Numbers

As mentioned, the table contains three repeats of each sample. The three repeats have to be assigned each to the same 'Sample'. In the table shown above, the 'Sample' column contains nothing but '1' for each spectrum. You can edit each cell directly by double-clicking on it, but there is a much easier way.

Click the 'Set Sample Numbers' button:



And this dialog will appear.

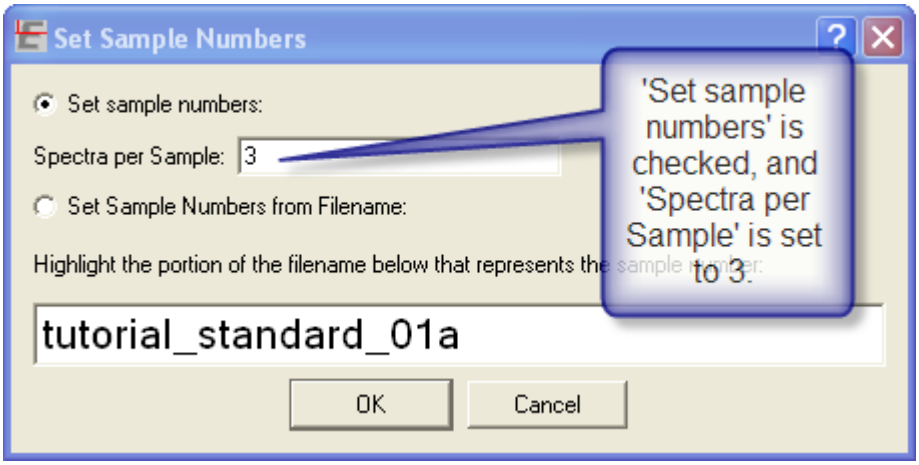


Select the radio button labeled ‘Set Sample Numbers from Filename’, then use the left mouse button to highlight the numeric portion of the filename that is displayed, in this case ‘01’. It’s important to include the leading ‘0’. Then click OK.

The spectra table will now look like this. Note how the ‘Sample’ column has been filled in correctly.

	1	2	3	4	5	6	7	8
1								
2	Data Set	Sample	Path	Filename	Subfile	Water	Methanol	Acetonitrile
3	calibration	1	C:/Documents and S	tutorial_standard ...	0	19.13	11.99	68.89
4	calibration	1	C:/Documents and S	tutorial_standard ...	0	19.13	11.99	68.89
5	calibration	1	C:/Documents and S	tutorial_standard ...	0	19.13	11.99	68.89
6	calibration	2	C:/Documents and S	tutorial_standard ...	0	18.72	25.46	55.82
7	calibration	2	C:/Documents and S	tutorial_standard ...	0	18.72	25.46	55.82
8	calibration	2	C:/Documents and S	tutorial_standard ...	0	18.72	25.46	55.82
9	calibration	3	C:/Documents and S	tutorial_standard ...	0	15.52	38.44	46.03
10	calibration	3	C:/Documents and S	tutorial_standard ...	0	15.52	38.44	46.03
11	calibration	3	C:/Documents and S	tutorial_standard ...	0	15.52	38.44	46.03
12	calibration	4	C:/Documents and S	tutorial_standard ...	0	15.61	49.56	34.83
13	calibration	4	C:/Documents and S	tutorial_standard ...	0	15.61	49.56	34.83
14	calibration	4	C:/Documents and S	tutorial_standard ...	0	15.61	49.56	34.83
15	calibration	5	C:/Documents and S	tutorial_standard ...	0	14.32	62	23.68
16	calibration	5	C:/Documents and S	tutorial_standard ...	0	14.32	62	23.68
17	calibration	5	C:/Documents and S	tutorial_standard ...	0	14.32	62	23.68

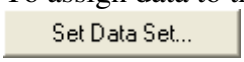
Note that in in this case, because there are the same number of repetitions for each sample, the ‘Set Sample Numbers’ could have been filled in this way and it would have had the same effect:



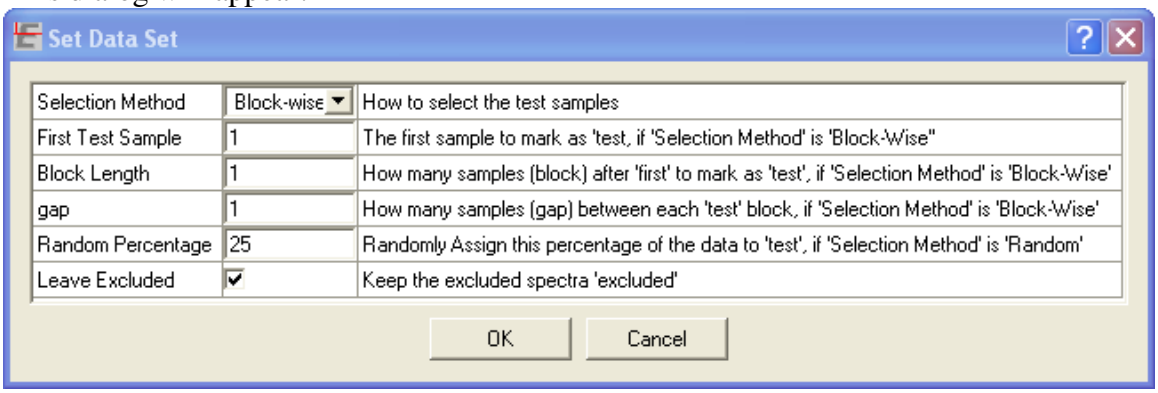
Setting the Data Set.

In a PLS method, it is common to split the data into calibration and test data sets. This is the reason that sample numbers for repeat samples are important, because you don’t want the same samples in both the calibration and test sets. Using a test set also speeds up validation, because cross-validation has to be performed if all the spectra belong to the calibration set, and cross-validation can be very time consuming for large data sets.

To assign data to the calibration and test sets, click the ‘Set Data Set’ button:



This dialog will appear:



Just click ‘OK’. The spectra table will now look like this:

	1	2	3	4	5	6	7	8
1								
2	Data Set	Sample	Path	Filename	Subfile	Water	Methanol	Acetonitrile
3	test	1	C:/Documents and S	tutorial_standard	0	19.13	11.99	68.89
4	test	1	C:/Documents and S	tutorial_standard	0	19.13	11.99	68.89
5	test	1	C:/Documents and S	tutorial_standard	0	19.13	11.99	68.89
6	calibration	2	C:/Documents and S	tutorial_standard	0	18.72	25.46	55.82
7	calibration	2	C:/Documents and S	tutorial_standard	0	18.72	25.46	55.82
8	calibration	2	C:/Documents and S	tutorial_standard	0	18.72	25.46	55.82
9	test	3	C:/Documents and S	tutorial_standard	0	15.52	38.44	46.03
10	test	3	C:/Documents and S	tutorial_standard	0	15.52	38.44	46.03
11	test	3	C:/Documents and S	tutorial_standard	0	15.52	38.44	46.03
12	calibration	4	C:/Documents and S	tutorial_standard	0	15.61	49.56	34.83
13	calibration	4	C:/Documents and S	tutorial_standard	0	15.61	49.56	34.83
14	calibration	4	C:/Documents and S	tutorial_standard	0	15.61	49.56	34.83
15	test	5	C:/Documents and S	tutorial_standard	0	14.32	62	23.68
16	test	5	C:/Documents and S	tutorial_standard	0	14.32	62	23.68
17	test	5	C:/Documents and S	tutorial_standard	0	14.32	62	23.68

Note how all three spectra for sample 1 are assigned to the ‘test’ set, the three spectra for sample 2 are in the ‘calibration’ set, and so on.

Entering the Concentration Information.

You can double-click the left mouse button on any of the yellow cells to directly enter the concentration information. Here is the table of concentrations:

Mixture #	% Water	% Methanol	% Acetonitrile
1	19.13	11.99	68.89
2	18.72	25.46	55.82
3	15.52	38.44	46.03
4	15.61	49.56	34.83
5	14.32	62.00	23.68
6	15.69	72.40	11.91

7	30.99	11.73	57.29
8	30.35	23.94	45.71
9	30.72	35.16	34.13
10	30.68	46.59	22.73
11	30.72	57.91	11.37
12	44.74	11.28	43.98
13	44.48	22.45	33.07
14	43.89	33.97	22.14
15	44.27	44.68	11.05
16	57.01	10.82	32.18
17	56.79	22.02	21.19
18	56.93	32.45	10.62
19	68.35	10.75	20.90
20	68.39	21.29	10.32
21	79.64	10.28	10.08

Note that there are 63 spectra; there are three repeats for each mixture.
Here is the same table, with the repeats added:

19.13	11.99	68.89
19.13	11.99	68.89
19.13	11.99	68.89
18.72	25.46	55.82
18.72	25.46	55.82
18.72	25.46	55.82
15.52	38.44	46.03
15.52	38.44	46.03
15.52	38.44	46.03
15.61	49.56	34.83
15.61	49.56	34.83
15.61	49.56	34.83
14.32	62.00	23.68
14.32	62.00	23.68
14.32	62.00	23.68
15.69	72.40	11.91
15.69	72.40	11.91
15.69	72.40	11.91
30.99	11.73	57.29
30.99	11.73	57.29
30.99	11.73	57.29
30.35	23.94	45.71
30.35	23.94	45.71
30.35	23.94	45.71
30.72	35.16	34.13
30.72	35.16	34.13

30.72	35.16	34.13
30.68	46.59	22.73
30.68	46.59	22.73
30.68	46.59	22.73
30.72	57.91	11.37
30.72	57.91	11.37
30.72	57.91	11.37
44.74	11.28	43.98
44.74	11.28	43.98
44.74	11.28	43.98
44.48	22.45	33.07
44.48	22.45	33.07
44.48	22.45	33.07
43.89	33.97	22.14
43.89	33.97	22.14
43.89	33.97	22.14
44.27	44.68	11.05
44.27	44.68	11.05
44.27	44.68	11.05
57.01	10.82	32.18
57.01	10.82	32.18
57.01	10.82	32.18
56.79	22.02	21.19
56.79	22.02	21.19
56.79	22.02	21.19
56.93	32.45	10.62
56.93	32.45	10.62
56.93	32.45	10.62
68.35	10.75	20.90
68.35	10.75	20.90
68.35	10.75	20.90
68.39	21.29	10.32
68.39	21.29	10.32
68.39	21.29	10.32
79.64	10.28	10.08
79.64	10.28	10.08
79.64	10.28	10.08

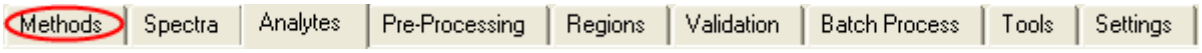
The easiest way to add the concentrations is to copy this table to the clipboard. This table is also installed in the tutorial directory as ‘concentrations.txt’. You can load concentrations.txt into any word processor and then copy the table to the windows clipboard. After the 63 lines of the complete concentration table are on the clipboard, right click on the first yellow cell in the spectra table in Essential FTIR, and choose ‘Paste’. The visible part of concentration table will now look like this:

6	7	8
Water	Methanol	Acetonitrile
19.13	11.99	68.89
19.13	11.99	68.89
19.13	11.99	68.89
18.72	25.46	55.82
18.72	25.46	55.82
18.72	25.46	55.82
15.52	38.44	46.03
15.52	38.44	46.03
15.52	38.44	46.03
15.61	49.56	34.83
15.61	49.56	34.83
15.61	49.56	34.83
14.32	62	23.68
14.32	62	23.68
14.32	62	23.68

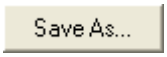
That was a lot easier than editing each cell individually.

Save the method.

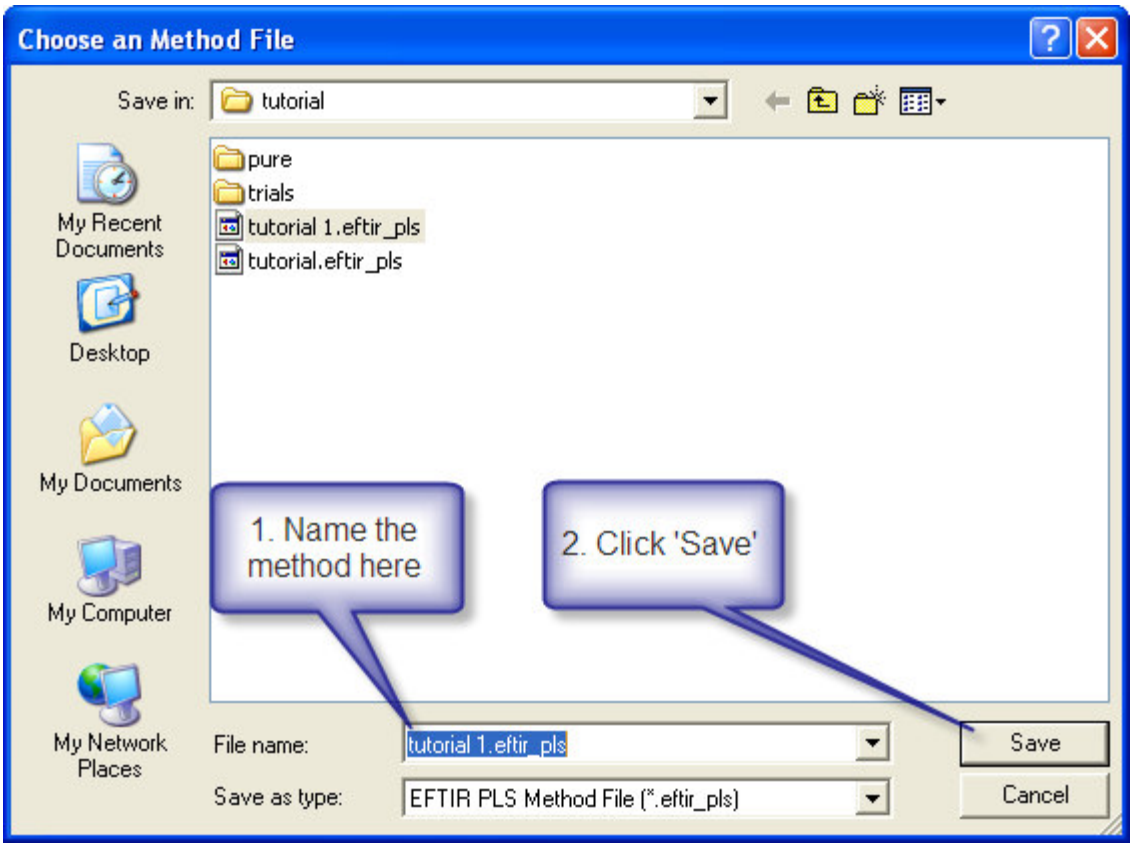
Click on the 'Methods' Tab:



And click the 'Save As…' button:

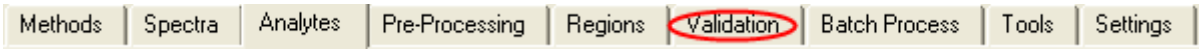


In the dialog that appears, type 'tutorial 1' and then click 'Save'

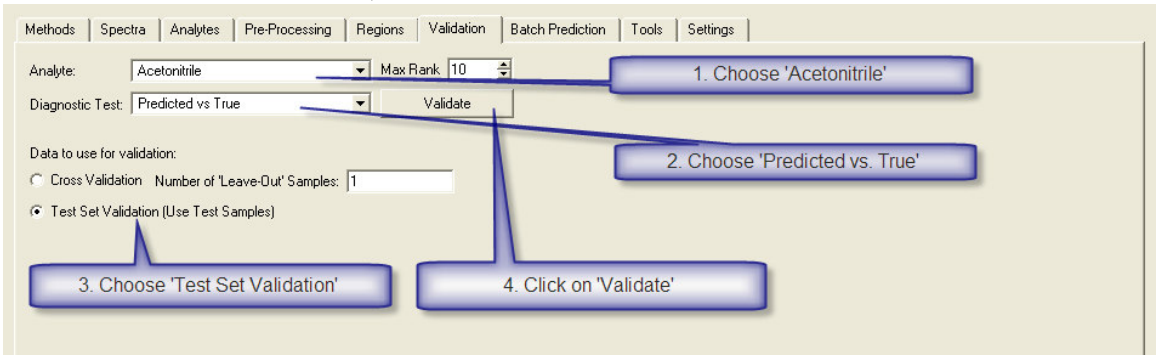


Testing and validating the method.

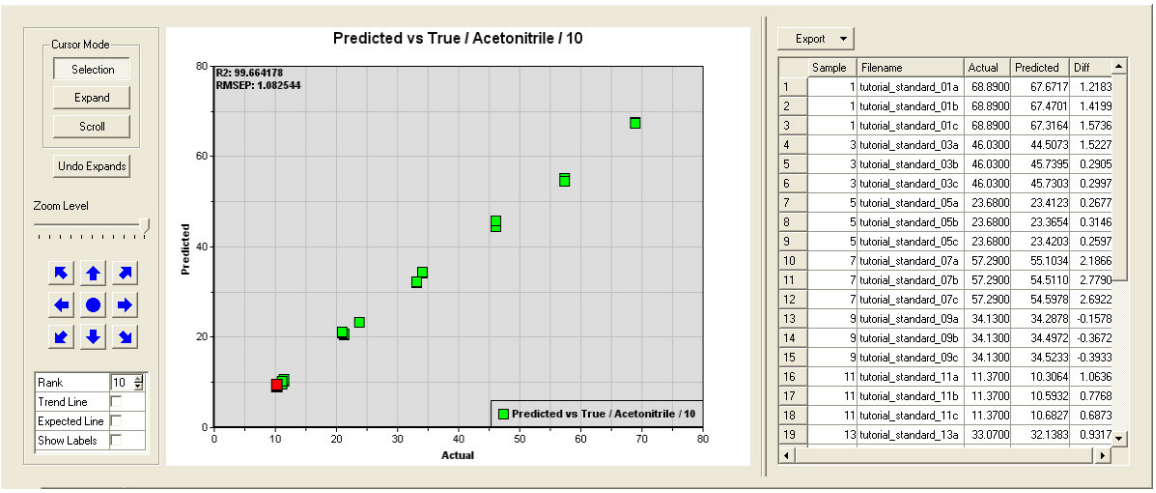
Click on the 'Validation' Tab



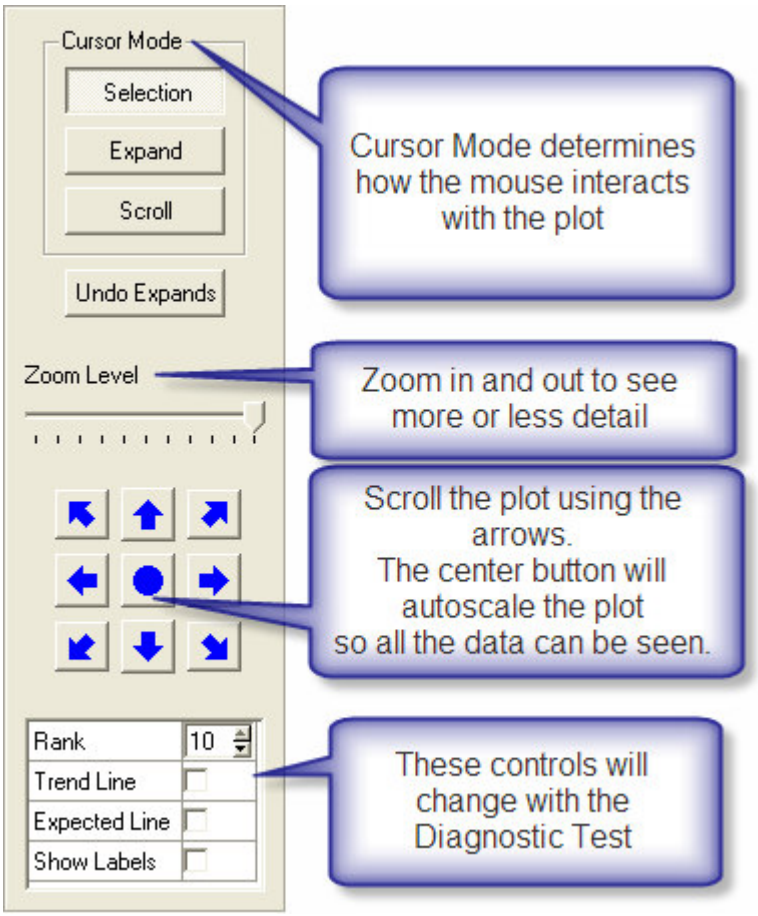
Choose 'Acetonitrile' as the Analyte, Choose 'Predicted vs. True' as the Diagnostic test, choose 'Test Set Validation', and then click 'Validate'



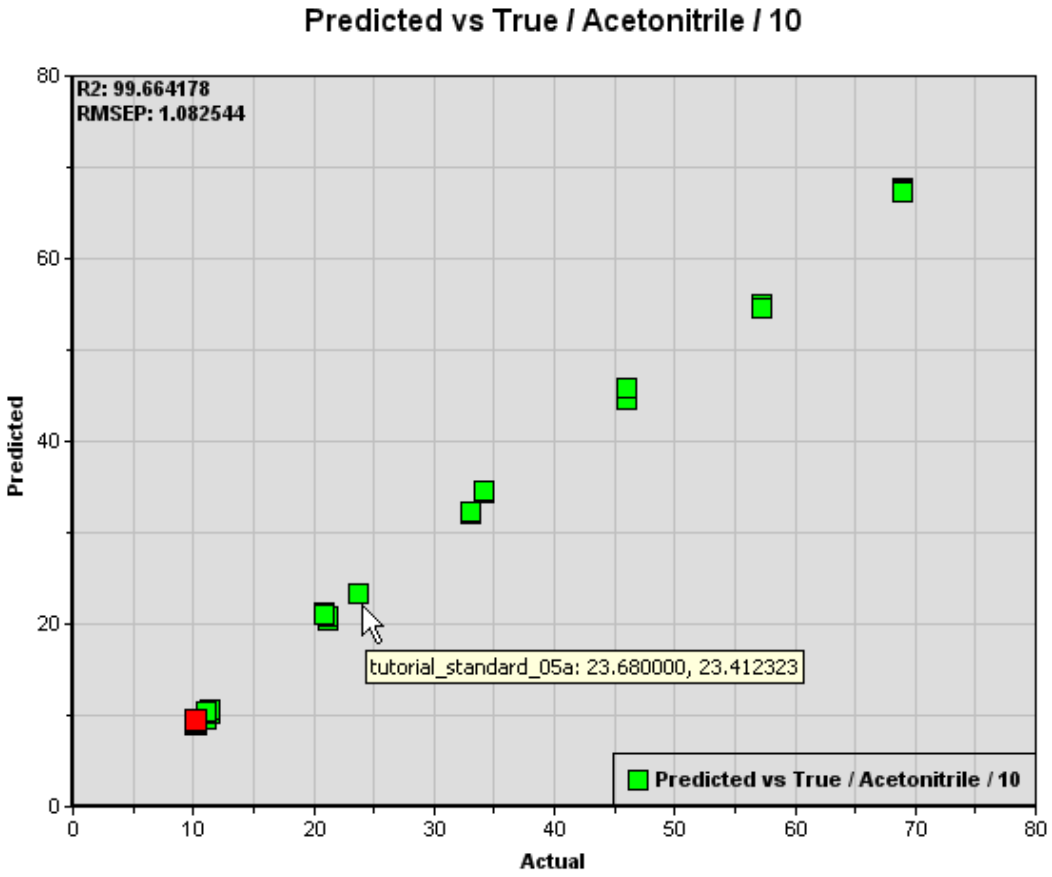
The method will be calibrated with the calibration samples from the Spectra table, and the method will be used to predict the concentration of Acetonitrile for the test samples. The resulting data will be plotted, as shown here:



There is a lot of information here.
In the center of the window is the plot of Predicted vs. True.
On the right is a table of the values that are in the plot. The 'Export' button can export the table to Excel, a disk file, or the Windows clipboard.
On the left are a variety of controls for handling the plot.

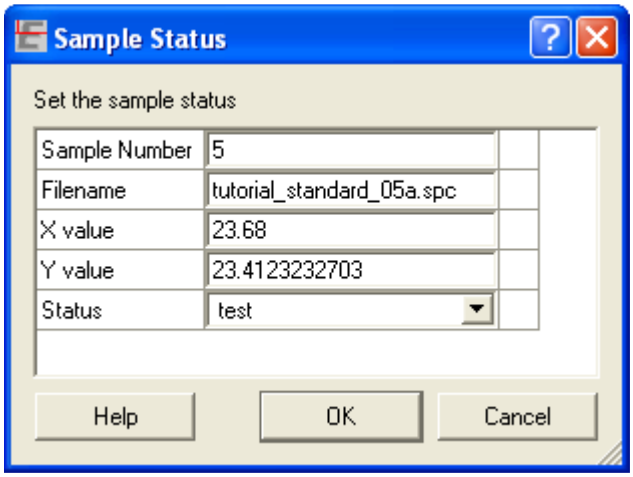


In the center plot, a pop-up 'Tool-Tip' will appear when the mouse hovers over a data point:



Note that some of the points are drawn in red, this means the software has flagged the sample as an outlier (more on this below).

On some of the diagnostic plots, if you left-click on a data point, a dialog will appear:

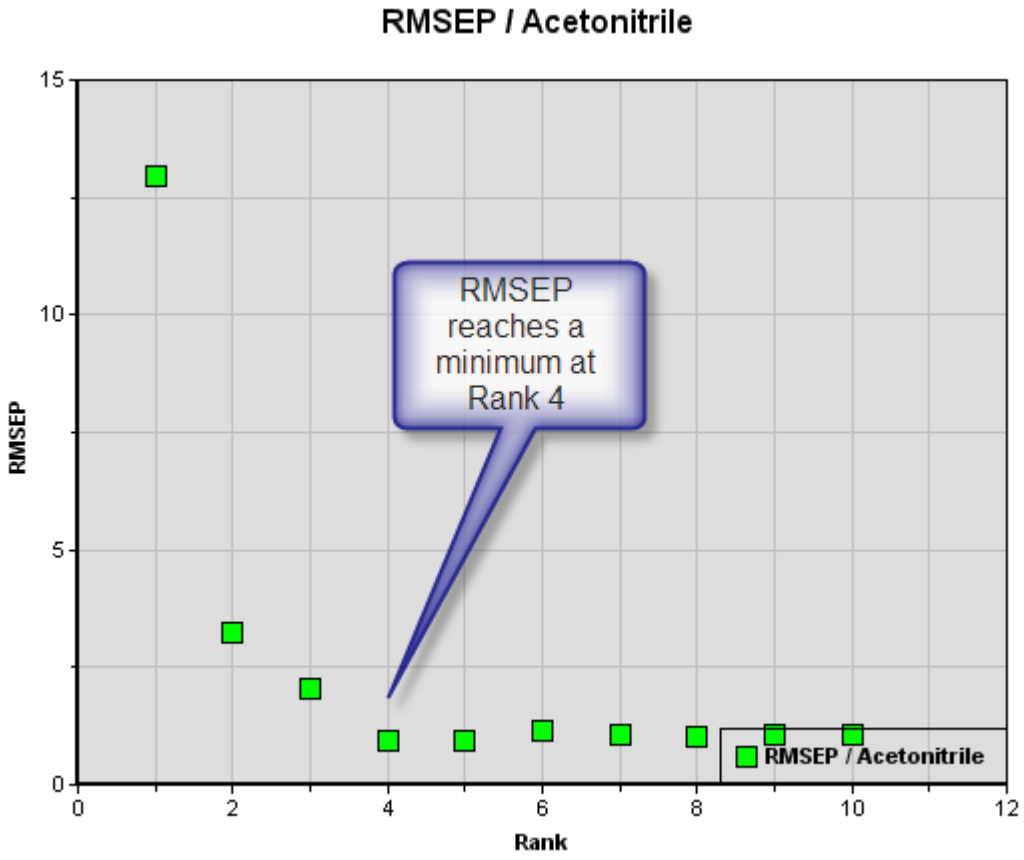


The important thing in the above dialog is the last line, labeled ‘Status’. This allows you to change the sample’s status between ‘test’, ‘calibration’ and ‘excluded’. This allows an interactive way to flag samples as outliers and exclude them from the calibration and test sets.

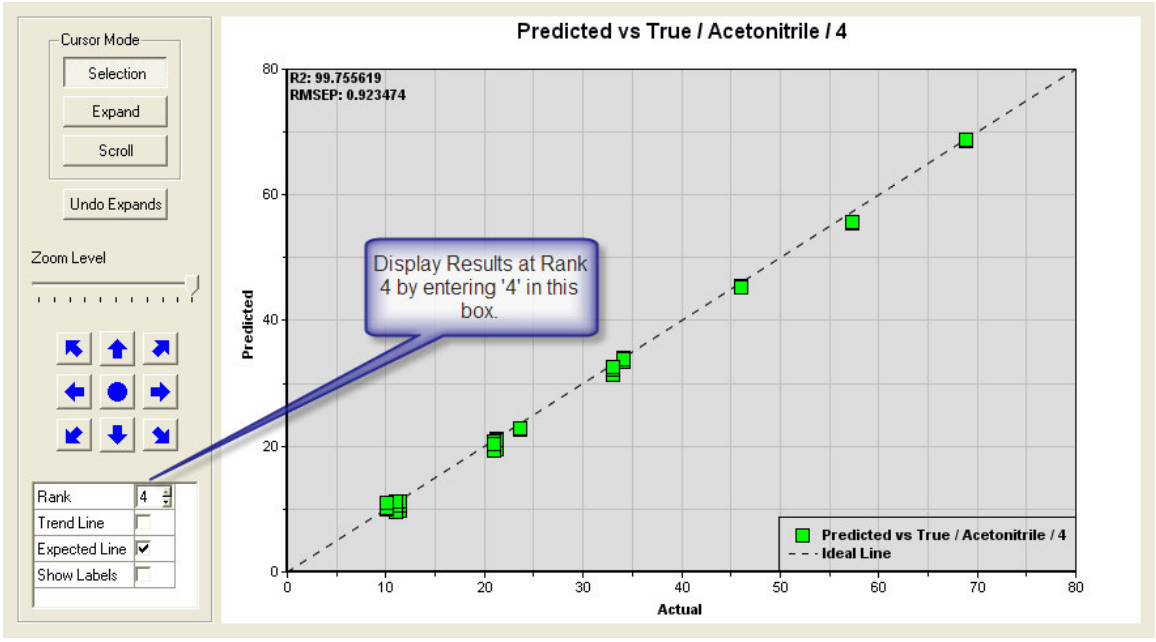
Outliers are draw in red, but in this particular case the red samples are not outliers. That’s because the data is over-fitted by the selection of 10 as the number of factors to use. For each analyte there is an optimal number of factors (also known as ‘rank’) to use for analysis.

Determining the Optimal Rank.

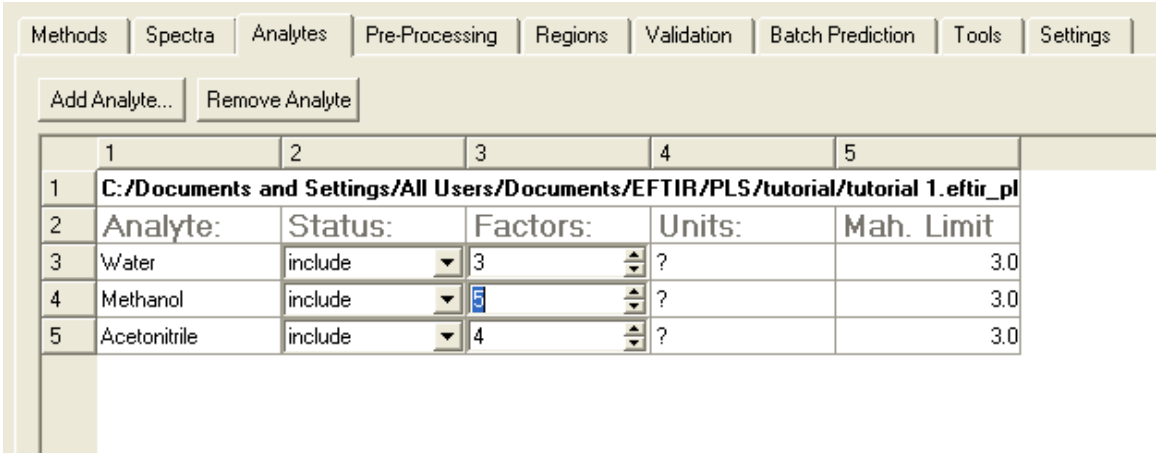
In the ‘Diagnostic Test’ list box, choose ‘Rank vs RMSE(CV,P). This will plot the ‘Root Mean Square of Prediction’ against rank, up to the number in the ‘Max Rank’ box. Here is what you should see:



It would appear that Rank 4 will give the best results for Acetonitrile. Returning to the ‘Predicted vs. True’ plot, now there are no outliers detected.



Doing the same analysis for Water and Methanol, the optimal number of factors (rank) for these was determined to be 3 and 5, respectively. On the 'Analytes' tab, enter these numbers in the Analyte Table:

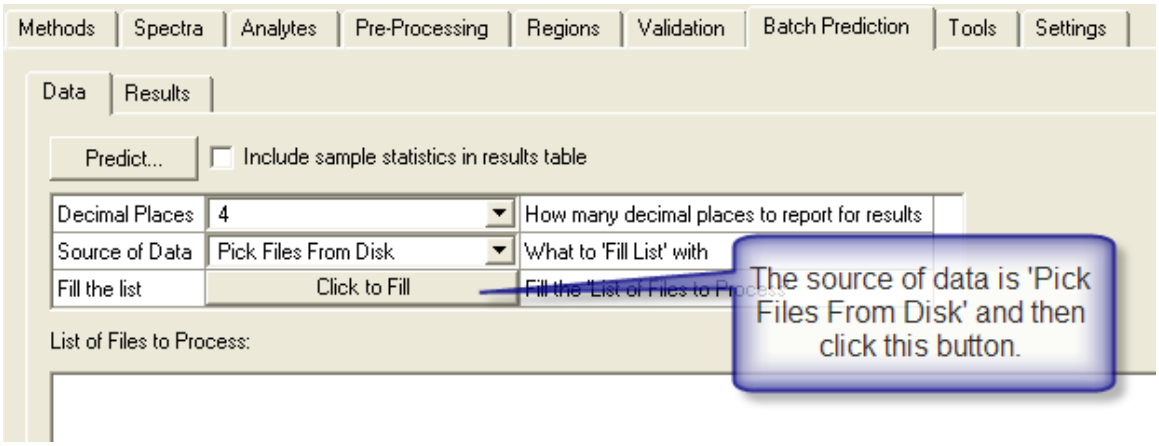


And save the method to disk by using the 'Save Method' button on the 'Methods' tab.

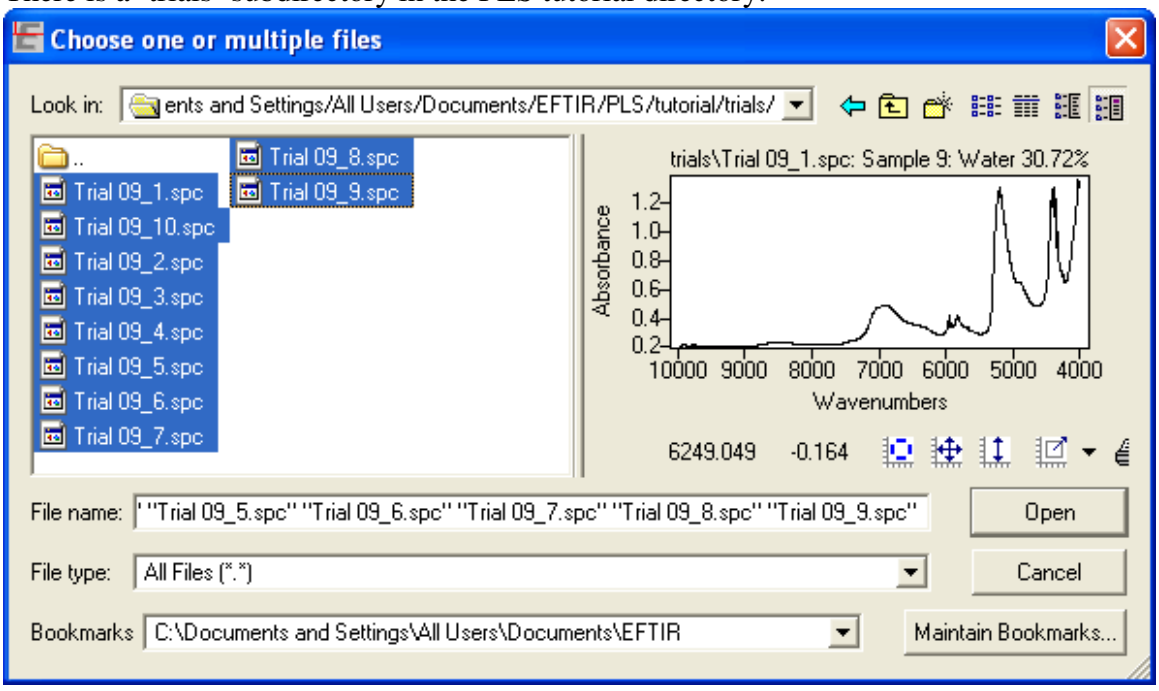
There is of course a lot more to say about testing and validating a PLS method. Please refer to the texts in the References section for in-depth discussions of this topic.

Batch Prediction

Now that we have a working method, we can apply it to some unknowns. Click on the 'Batch Prediction' tab and pick some files to analyze.



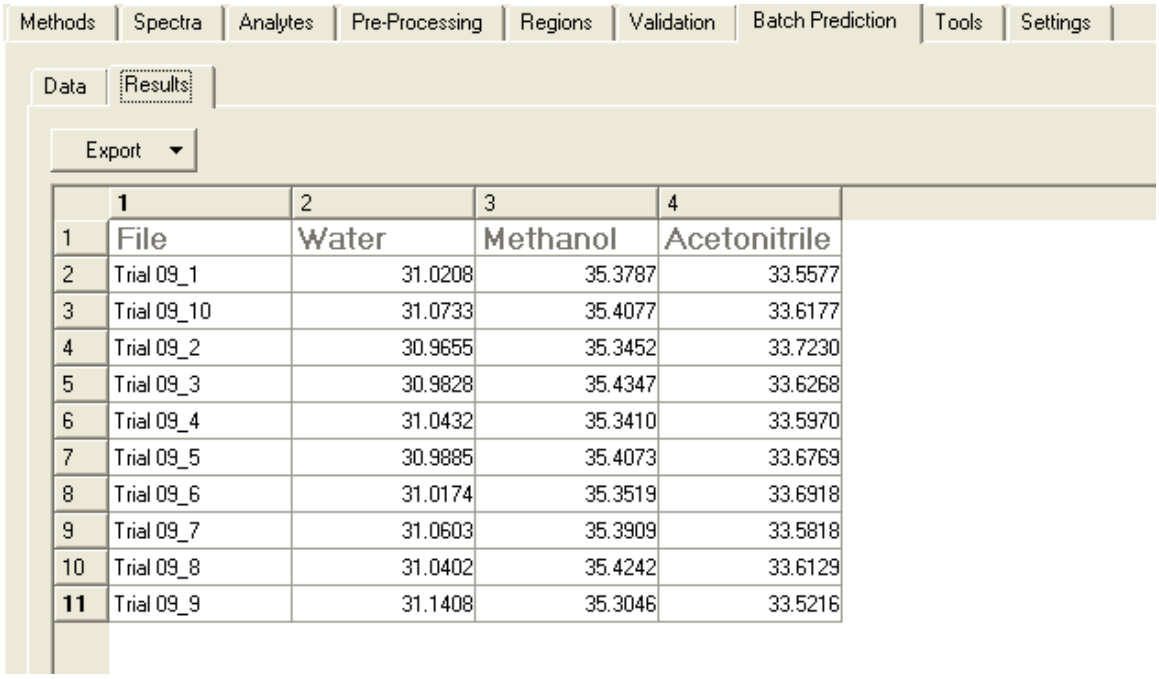
There is a 'trials' subdirectory in the PLS tutorial directory:



Choose all nine files and click 'Open'. The 'List of Files To Process' will be filled with the names of these files. These 9 files are repeats of the same mixture, containing these percentages of the analytes:

% Water	% Methanol	% Acetonitrile
30.72	35.16	34.13

Then click the ‘Predict’ button and the results table will be filled with the calculated concentrations:



	1	2	3	4
1	File	Water	Methanol	Acetonitrile
2	Trial 09_1	31.0208	35.3787	33.5577
3	Trial 09_10	31.0733	35.4077	33.6177
4	Trial 09_2	30.9655	35.3452	33.7230
5	Trial 09_3	30.9828	35.4347	33.6268
6	Trial 09_4	31.0432	35.3410	33.5970
7	Trial 09_5	30.9885	35.4073	33.6769
8	Trial 09_6	31.0174	35.3519	33.6918
9	Trial 09_7	31.0603	35.3909	33.5818
10	Trial 09_8	31.0402	35.4242	33.6129
11	Trial 09_9	31.1408	35.3046	33.5216

References

D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193.
H. Martens, T. Naes, *Multivariate Calibration*, J. Wiley & Sons(1989)
PLSplus/IQ User's Guide, Galactic Industries (2000)
Jorg-Peter Conzen, *Mutivariate Calibration*, Bruker Optik GmbH (2003)
H. Mark, J. Workman, *Statistics in Spectroscopy*, Academic Press (1991)
R Brereton*, Chemometrics Data Analysis for the Laboratory and Chemical Plant*, John Wiley & Sons (2003)
K Beebe, R. Pell, M. Seasholtz, *Chemometrics A Practical Guide*, John Wiley & Sons (1998)
T. Naes, T. Isaksson, T. Fearn, T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification*, NIR Publications (2002)