

Tutorial for the correlation spectroscopy function corrspec

This tutorial will describe the principles of the program `corrspec`. For more mathematical details see reference 1. For this tutorial familiarity with the pure variable approach as used in the function `purity` and described in the Self-Modeling Mixture Analysis section of the PLS_Toolbox manual. The approach used in `corrspec` aim at studying correlation maps of two spectroscopies. In this way it is possible, for example, to establish which peaks in an mid infra red (mid-IR) spectra are correlated with peaks in near infra red (near-IR) spectra. The figures used in this tutorial are generally the same ones as created in the demo of `corrspec`.

We will study the two data sets described in (1). The data sets are shown in Figure 1. The identical samples, a four component system of aromatic compounds, were analyzed by mid-IR and near-IR. For process monitoring near-IR is the method of choice, for spectral interpretation mid-IR is the method of choice. When relations are established between the two spectroscopies it is possible to get a better understanding of near-IR process monitoring data in terms of spectral interpretation. Establishing relations can be done by correlating the correlations between the wave numbers of the mid-IR data and the near-IR data. The correlation matrix can be plotted as a contour map, which is shown in Figure 2a.

The map only shows the positive correlations. The color bar at the right of the Figure indicates the relation between the colors and the correlation values in the correlation map. The map is very complicated and it is difficult to establish relations between the two spectroscopies. In order to make the map simpler we will give a lower weight to variables with a low value, so that variables in the noise range get a much lower weight. The weight factor needs to be set by the user, so some mathematical details need to be given in order to understand its effect. The weight factor is calculated as follows:

$$w_i = \frac{\mu_i}{\mu_i + \alpha} \quad (1)$$

where w_i is the weight factor for variable i (e.g. a wavenumber in the mid-IR data) and μ_i the mean value of variable i. The so-called offset α , with a typical value of 1-3% of the maximum of all the mean values, takes care that w_i for variables with a low mean value gets a low value, while variables with a high mean value get a value close to one. So if the values of a wave number in the mid-IR spectra have a correlation with the values in the near-IR spectra of, for example, 0.8, and the weight factor of the mid-IR is 0.6 and the weight factor of the near-IR is 0.7, the final correlation will be $0.6 \times 0.7 \times 0.8$.

The maximum weight factor using Equation 1 is less than one. A correction factor has been added so that the maximum weight factor is one (not shown).

The effect of applying the typical offset of 3 for both spectroscopies (resulting in values of α for each of the spectroscopies of 3% of the maximum mean values) results in the correlation map in Figure 2b.

The map is much simpler than the one in Figure 2a, but still complex. The reason is that there are several causes for correlations. For example, typical peaks for a component, say component A, would be correlated within each spectroscopy and also between the spectroscopies. The peaks for component B would be correlated, but the peaks of component A and component B would not necessarily be correlated with each other. It would be nice if the correlation map could be split into simpler correlation maps, in this case, correlation maps of the underlying 4 components. This is what the function `corrspc` does. We will now run the `corrspc` function with the 2 data sets discussed above using the following commands:

```
>> load data_mid_IR  
>> load data_near_IR  
>> corrspc(data_mid_IR,data_near_IR,4);
```

The function `corrspc` has a default value for the offset of 3 (i.e., 3% of the maximum mean value). The third argument indicates the number of components we will resolve.

The function `corrspc` creates Figure 3a, which we have seen before in Figure 2b, but now there is a cursor in the plot. The cursor position is determined as follows:

- a) The variables (about 1350 cm^{-1} for the mid-IR and about 6000 cm^{-1} for the near-IR) are correlated with each other and
- b) The variables are pure for the same component (see function `purity`).

Thus variable pair $1350/6000\text{ cm}^{-1}$ has the highest so called co-purity in the data set (1). When the term variable pair is used the spectroscopy on the x-axis (mid-IR) will be used as the first one, and the spectroscopy on the y-axis (near-IR) will be used as the second one.

A pure variable for a certain component has contributions from only one of the components in the mixture. As such, its intensities are proportional to the actual concentrations and can be used to resolve a data set into pure component spectra and their contributions. The term "contributions" is used instead of concentrations since there are unknown factors between the actual concentrations and the values we calculate, which are also different for each component. In order to indicate this, the term contribution is used.

Summarizing the cursor indicates a variable pair typical for a certain component, for both spectroscopies. In order to further simplify the correlation map the next step is to take out that part of the correlation map that is correlated with the variable pair $1350/6000\text{ cm}^{-1}$. In other words, take out that part the correlation map based on the relation with the variable pair. This is achieved by clicking the mouse, while the mouse cursor is in the Figure. This results in Figure 3b, which obviously is much simpler than Figure 2a. The local maximum at the variable pair $1350/6000\text{ cm}^{-1}$ disappeared completely, as expected. Other

parts of the correlation map also disappeared, because of a high correlation with the variable pair. The parts of the correlation map that have no (or a low) correlation with the variable pair of the first component stand out now. In order to get a better understanding of correlation spectroscopy we will reproduce Figure 3b in Figure 4 in a larger format.

When we study Figure 4 closely, we see the following

- a) At point **1** we have the variable pair $1600/6700 \text{ cm}^{-1}$ with a high co-purity.
- b) There is a series of high correlations along the line between **1** and **2**. These wave numbers are highly correlated with the variable pair $1600/6700 \text{ cm}^{-1}$. This indicates that the peak that is pure for the second mid-IR component (1600 cm^{-1} of the variable pair $1600/6700 \text{ cm}^{-1}$), is highly correlated with all the other peaks on the line between **1** and **2** and the maxima are likely to be also typical for the second component of the mid-IR spectra, which we will be confirmed later.
- c) Similarly, the peaks 6700 cm^{-1} and 5000 cm^{-1} , which lie on the line between **1** and **3**, are most likely wave numbers typical for the second near-IR component.
- d) The near-IR 5000 cm^{-1} at point **3** is correlated with the near-IR 6700 cm^{-1} at point **1** and the mid-IR 850 cm^{-1} at point **2** is correlated with mid-IR 1600 cm^{-1} at point **1**, which leads to the conclusion that the mid-IR 850 cm^{-1} should have a high correlation with near-IR 5000 cm^{-1} , which is indeed indicated by a high correlation at **4**.

In this way we can draw rectangles of correlated peaks, as is done in Figure 4. When we correct for this pure variable set we should take out all the points with a high correlation, which is indeed the case, see Figure 3c.

The highly correlated wave numbers in Figure 3c are all connected by the vertical part of the crosshair cursor. This is a likely indication that for the mid-IR spectrum for the third component there is a single peak at around 900 cm^{-1} and a series of peaks for the near-IR spectra at 5100 cm^{-1} , 5200 cm^{-1} and 6500 cm^{-1} , which we will see confirmed later.

We also see a series of high correlations along the horizontal line at about 6000 cm^{-1} which are not correlated with the point defined by the cross hair cursor, since we do not see a high correlation at 6000 cm^{-1} on the vertical part of the cross hair. As a consequence, this must be a separate component, which will not disappear after clicking the mouse.

After taking away this component by a mouse click we obtain Figure 3d, which indeed leaves the high correlations along the horizontal line at about 6000 cm^{-1} . We can connect most of the remaining points in rectangles as shown in Figure 4, so we should not have much left after taking out this component, which is indeed the case, see Figure 3e.

Figure 3e still seems shows some structure, but we see that the crosshair cursor is not on any of these maxima, which is an indication that this correlation map is based on noise.

At this point **4** pure variable sets were determined. The next step is to use the pure variable intensities of the mid-IR data set to resolve the mid-IR data into the pure components (see Figure 5a) and their contributions (see Figure 5b) and is to use the pure

variable intensities of the near-IR data set to resolve the near-IR data into the pure components (see Figure 6a) and their contributions (see Figure 6b). This is achieved by a mouse click. The resolved components match well with the spectra of the separately analyzed components (1).

At this point we can confirm the points raised above when discussing Figure 4. The second resolved component shown in Figure 5a indeed shows a series of peaks we see on the line **1-2** in Figure 4. Similarly, we see the peaks of the second component in the near-IR spectrum in Figure 6a on line **1-2** in Figure 4. Similarly, we recognized that there would be one major peak for the third resolved mid-IR spectrum in Figure 5a from Figure 3c on the horizontal part of the crosshair cursor, and a series of peaks for the third resolved near-IR spectrum shown in Figure 6a from the vertical part of the crosshair cursor in Figure 3.

Since the pure variables sets were based on both purity and correlations we expect the contributions to be correlated. Therefore, the contributions of the mid-IR data set are plotted versus the intensities of the near-IR data set. A reasonable correlation can be observed, see Figure 7 (click on mouse for to generate this Figure and also the next Figures).

From the resolved results, correlation maps are calculated for each of the four resolved components, see Figure 8. Figure 8a shows the original correlation map, shown before in Figure 3a. From the resolved results shown in Figures 5 and 6 the data sets can be reconstructed. If the proper number of components has been resolved, the correlation map calculated from the reconstructed data sets, shown in Figure 8b, should be very similar to the original correlation map in Figure 8a, which is indeed the case here.

In Figure 9c we see the sum of the resolved maps. Although one would expect this map to be similar to Figure 8b this is not the case because the correlations between different components are not present in the resolved maps. A mathematical explanation is given in (1). An overlay plot of the four resolved components, each with a different color, is given in Figure 8d. A plot of the individual resolved maps is given in Figure 9.

This concludes the construction of the model that established the relation between the mid-IR data set and the near-IR data set. With this model we can now predict a near-IR spectrum from a mid-IR spectrum and vice versa. We do not have an unknown spectrum to check the validity of the prediction, so we will calculate a model after leaving out one of the samples, sample 8, from both data sets using the following commands:

```
>> data_mid_IR_reduced=data_mid_IR([1:7,9:end]);
>> data_near_IR_reduced=data_near_IR([1:7,9:end]);
```

Next, we will put the spectra we left out in separate arrays:

```
>> mid_IR2predict=data_mid_IR(8,:);
>> near_IR2predict=data_near_IR(8,:);
```

We will now build a model of 4 components (going through the same process as described above again).

```
>> model = corrspec(data_mid_IR_reduced,data_near_IR_reduced,4);
```

With this model we can now predict a mid-IR spectrum from a near-IR spectrum as follows:

```
>> model2=corrspec([],near_IR2predict,model);
```

A plot of the results shows the similarity between the predicted spectrum and the actual spectrum, see Figure 10a.

```
>> plot(data_mid_IR.axissscale{2},model2.loads{2},...
    data_mid_IR.axissscale{2},mid_IR2predict.data)
legend('mid-IR predicted from near-IR','actual mid-IR');
```

Similarly, we can predict the near-IR spectrum from the mid-IR spectrum, see Figure 10b.

```
>> model2=corrspec(mid_IR2predict,[],model);
```

Plot as follows:

```
>> plot(data_near_IR.axissscale{2},model2.loads{4},...
    data_near_IR.axissscale{2},near_IR2predict.data)
legend('mid-IR predicted from near-IR','actual mid-IR');
```

Both predictions can be made with one command:

```
>> model2=corrspec(mid_IR2predict,near_IR2predict,model);
```

In conclusion, the task of correlation spectroscopy is to establish relations between spectral variables. We have seen that correlation maps can be very complicated, which makes it difficult to establish these relations. The function `corrspec` enables us to simplify correlation maps which makes it easier to establish the relations between the spectral variables. This is achieved as follows:

- Giving lower intensity variables (noise range) a lower contributions in the correlation maps through the use of an offset.
- Resolving the correlation maps of mixture data sets into correlation maps of the individual components and their associated spectra.

References

- 1) W. Windig, D.E. Margevich, W.P. McKenna,
A novel tool for two-dimensional (2D) correlation spectroscopy,
Chemometrics and Intelligent Laboratory Systems, 28, 1995, 108-128

Figure 1. The mid-IR spectra in (a) and the near IR spectra of a series of mixtures of aromatic compounds in (b).

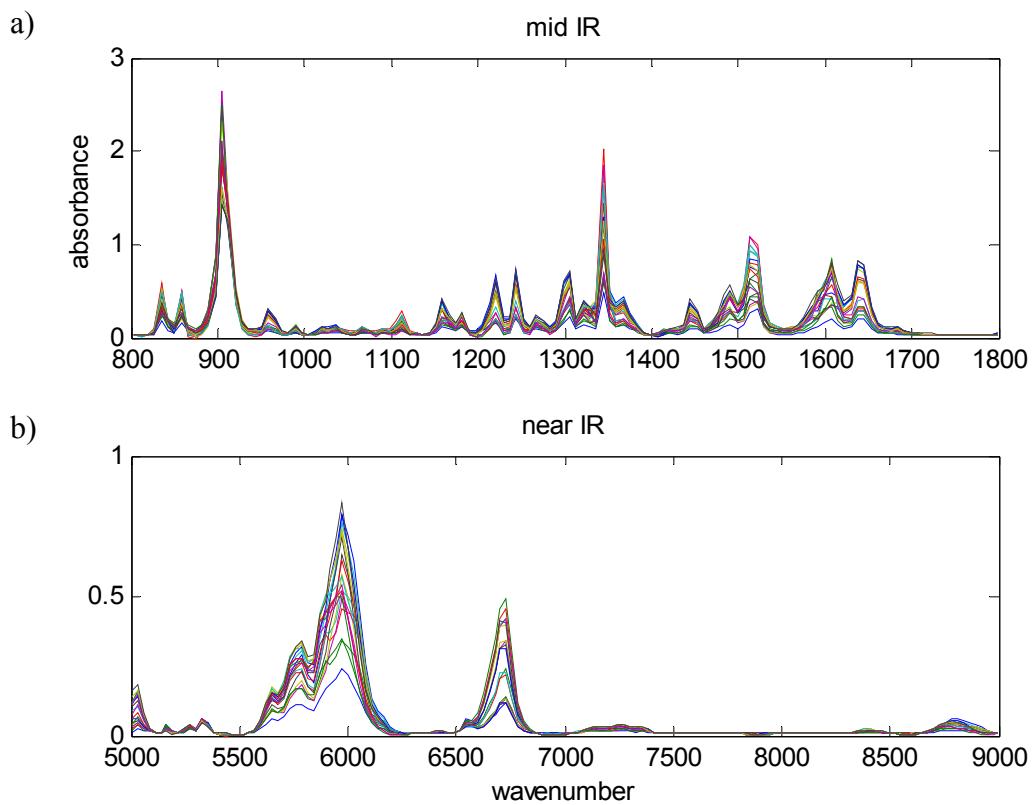


Figure 2. Correlation map of mid-IR and near-IR. In (a) the normally used correlations are displayed, in (b) the variables have been weighted according to their mean intensity (eq. 1), which results in minimizing the influence of low intensity (noise level) variables.

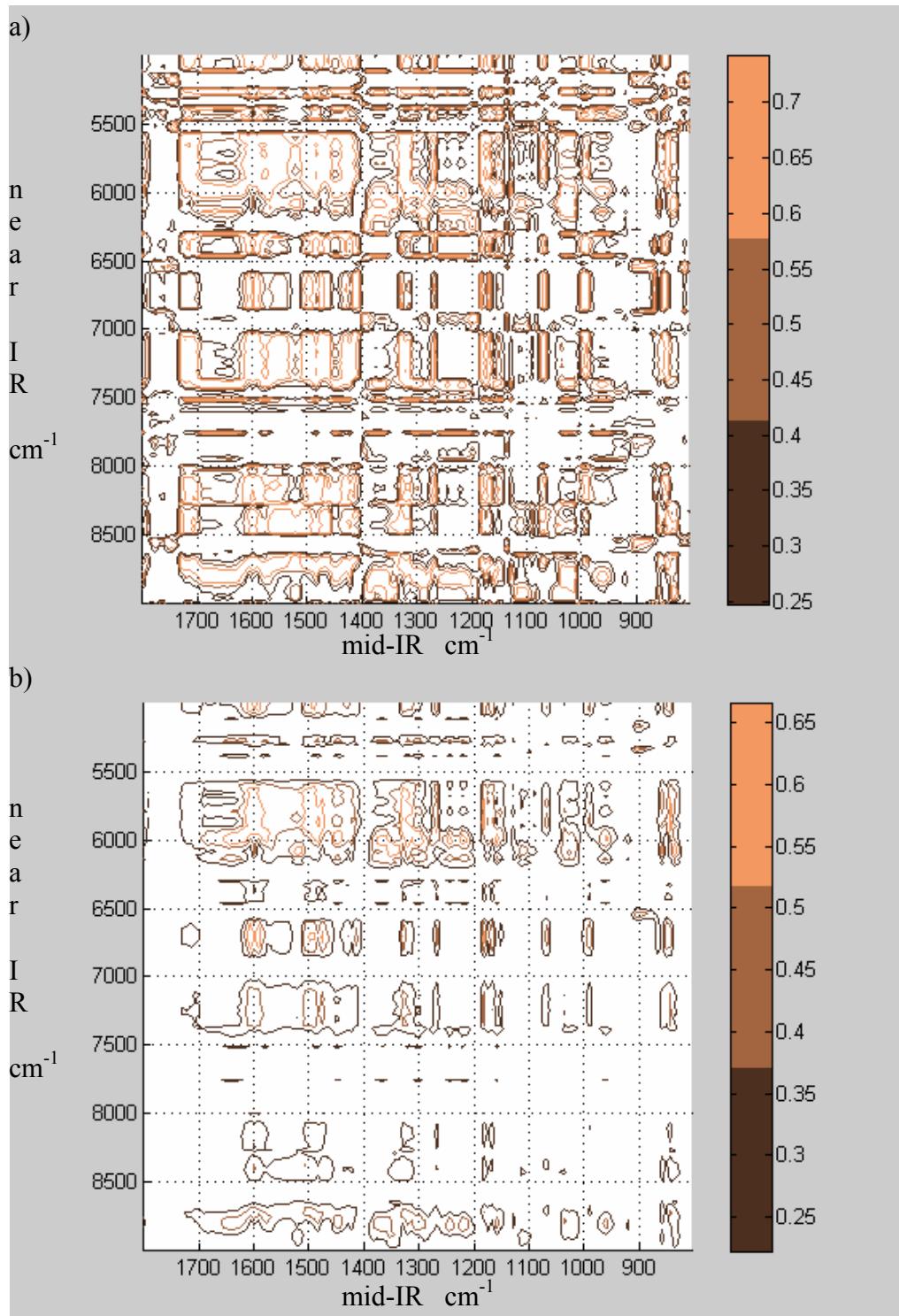


Figure 3. The successive series of correlation maps generated by the program corrspec: (a) shows the first correlation map, with an offset of 3; (b) shows the second correlation map, after taking out the information related with the first variable pair. Taking out the information related to the second pure variabe pair (c) results. The next step results in (d). In (e) we see that the cursor is not on a local maximum, indicating noise. This confirms the presence of 4 components.

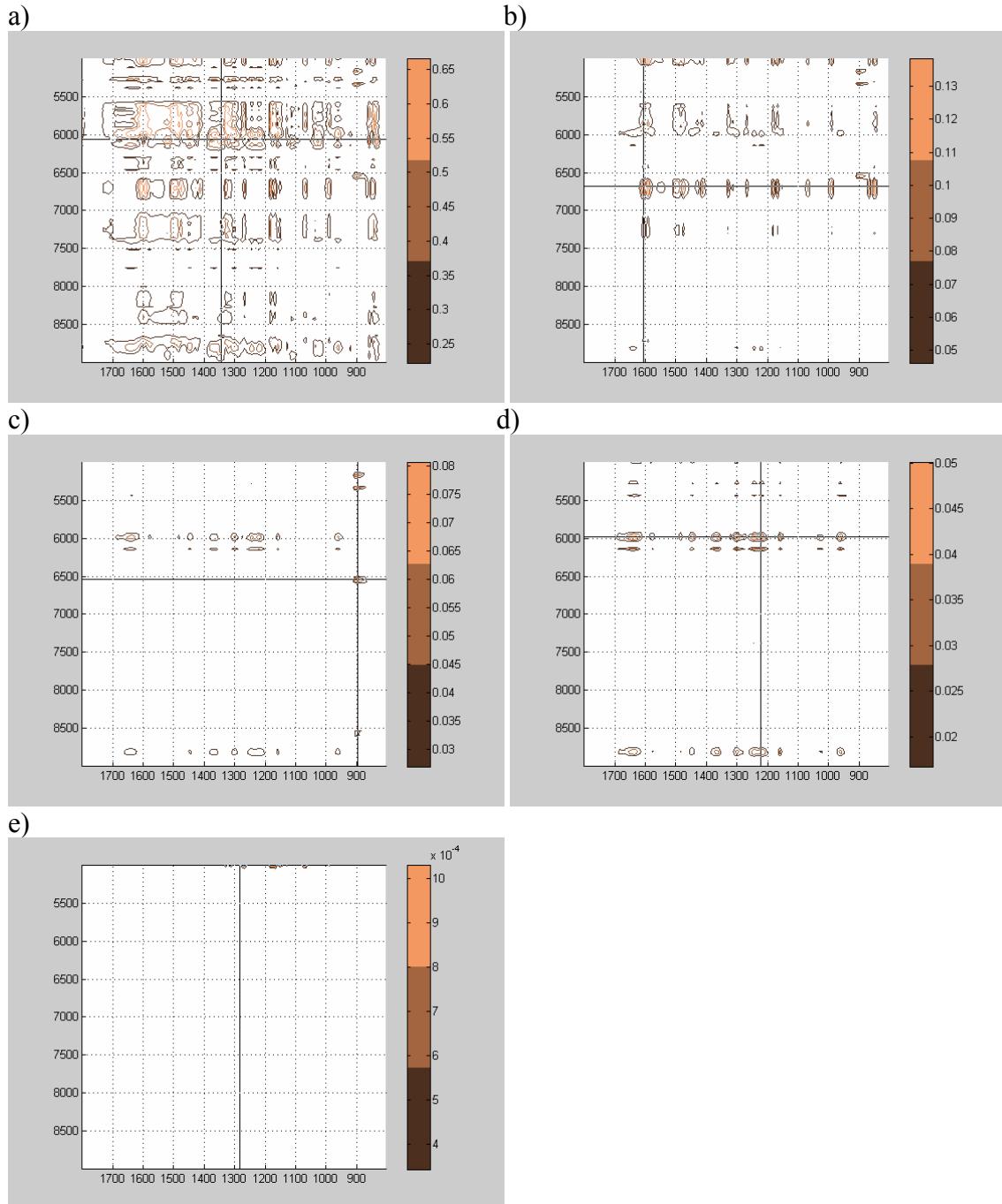


Figure 4. An enlarged version of Figure 3b with highly correlated features interconnected.

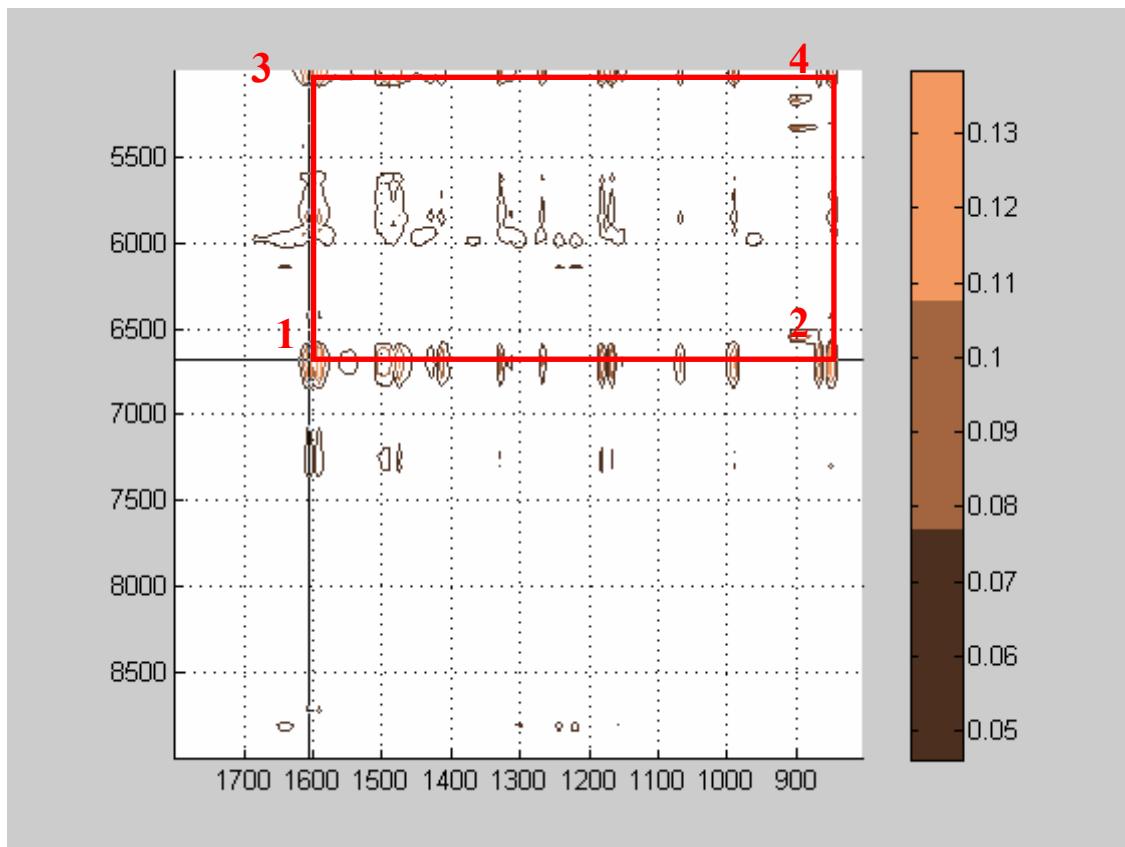


Figure 5. The four resolved component spectra of the mid-IR data set are shown in (a). The associated contributions are shown in (b).

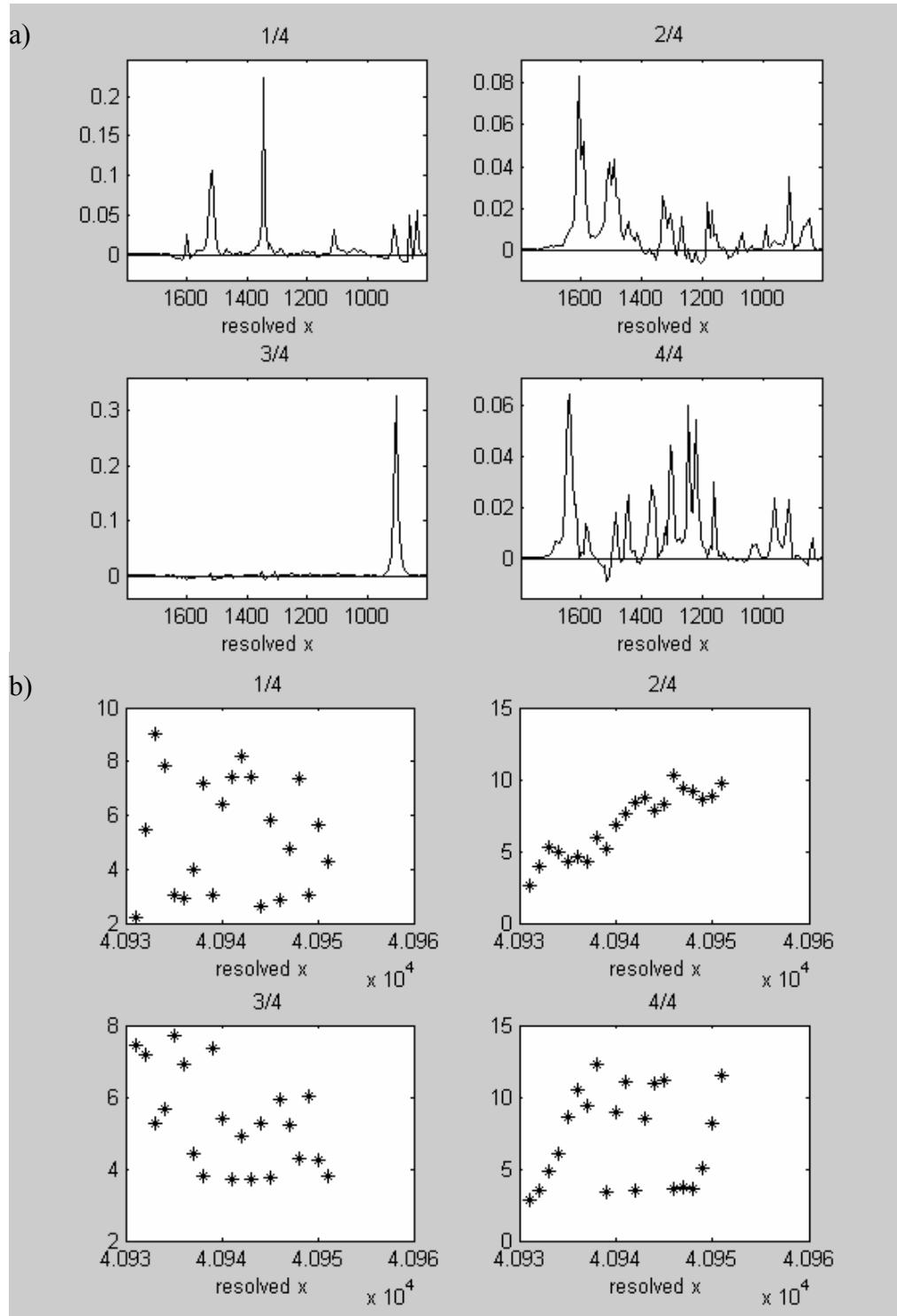


Figure 6. The four resolved component spectra of the mid-IR data set are shown in (a). The associated contributions are shown in (b).

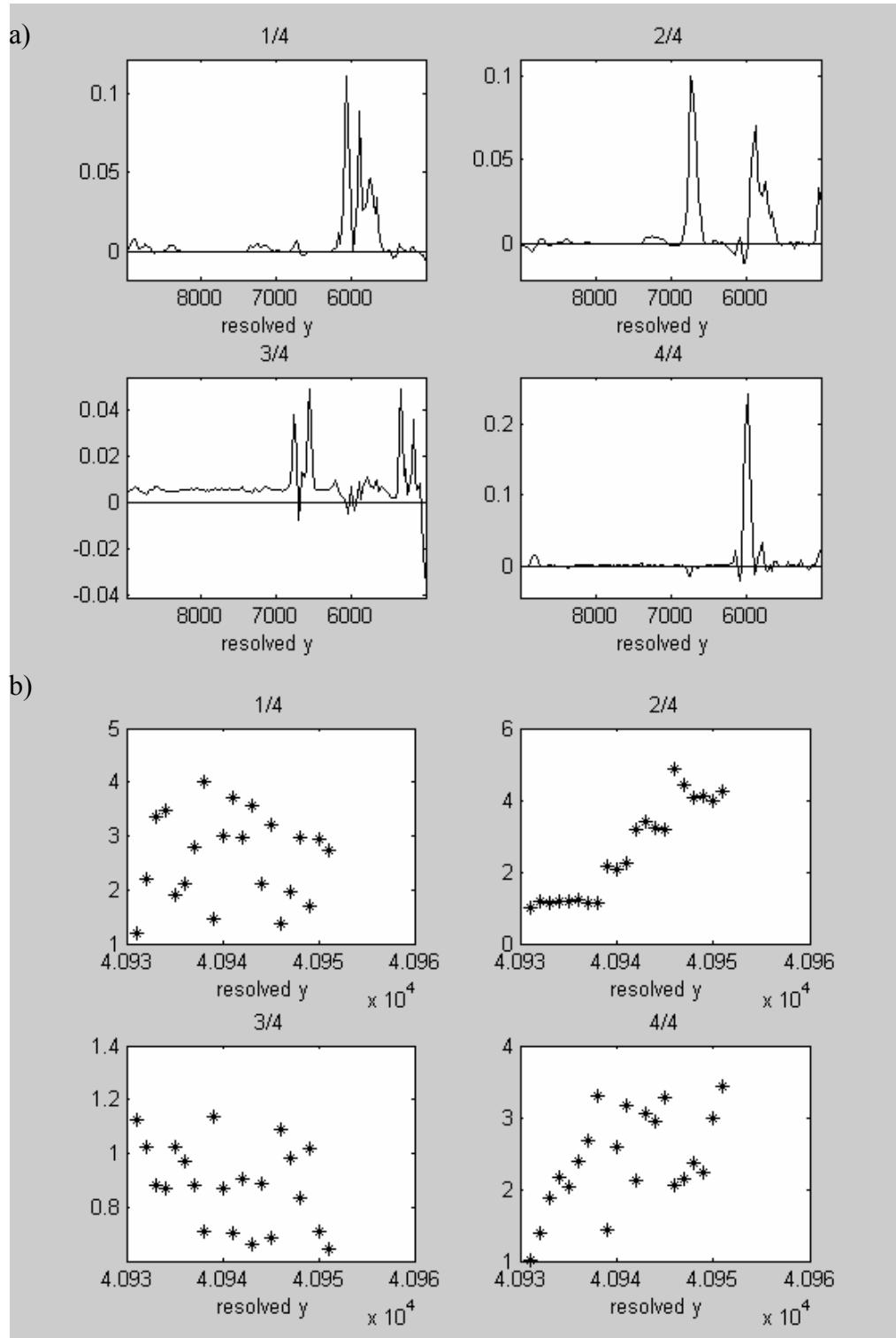


Figure 7. The resolved contributions of the mid-IR data set (shown in Figure 5b) versus the resolved contributions of the near-IR data set (shown in Figure 6b). A reasonable correlation can be observed.

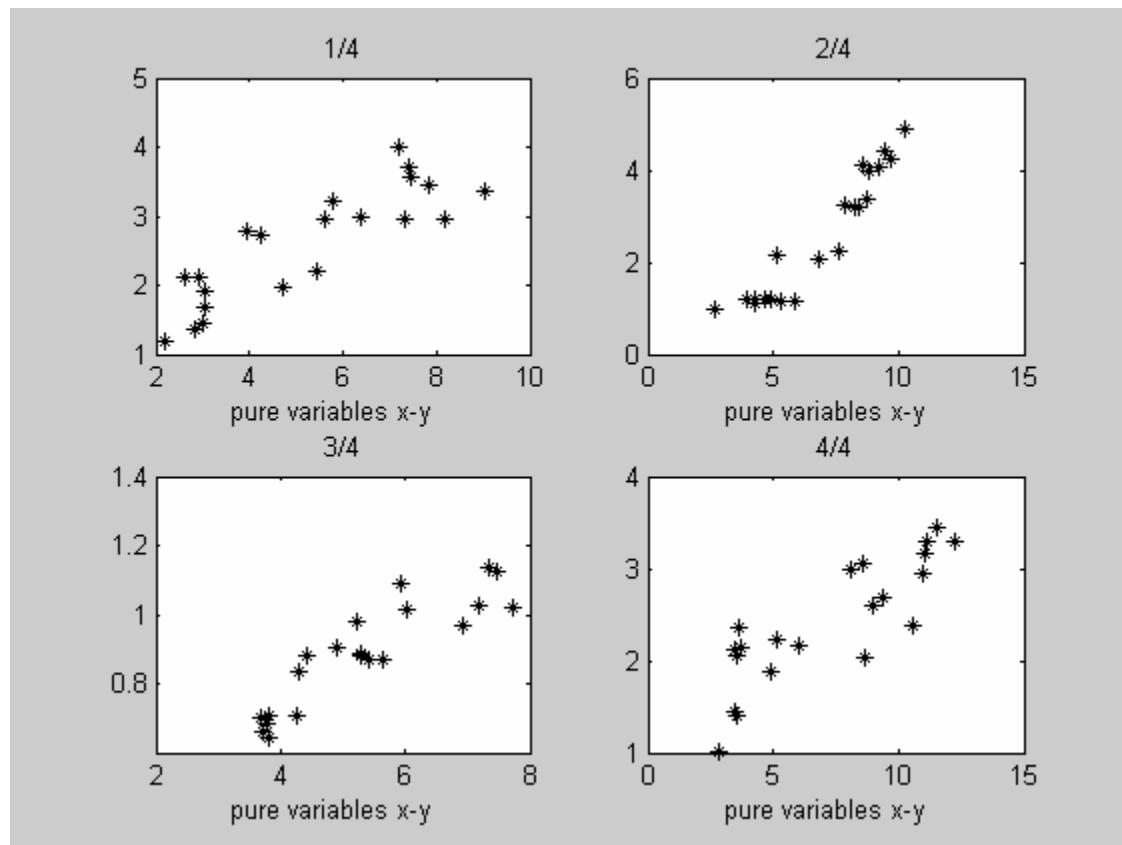


Figure 8. In (a) we see the original correlation map. In (b) we see the correlation map based on the reconstructed data files. In (c) we see the sum of the resolved matrices, which is similar to (b) but not the same for mathematical reasons. In (d) we see an overlay of the four resolved correlation maps, each with a different color.

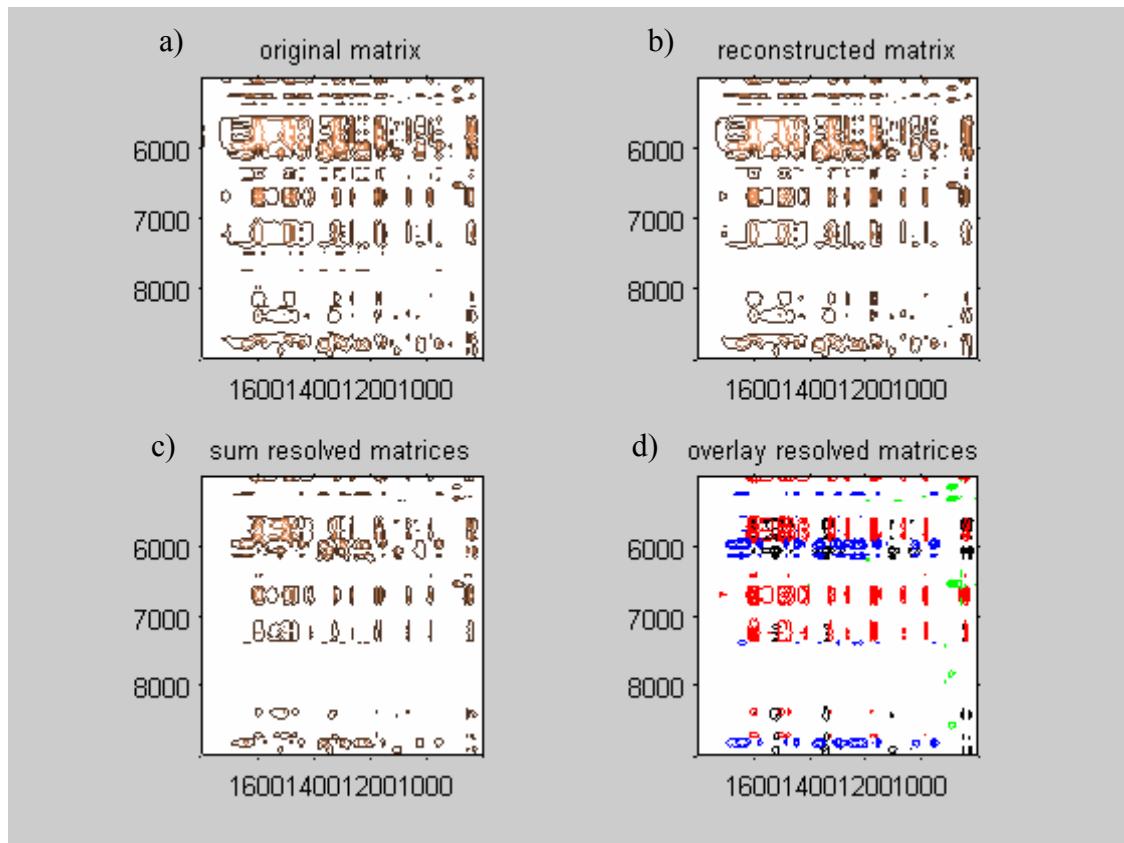


Figure 9. The four resolved correlation maps.

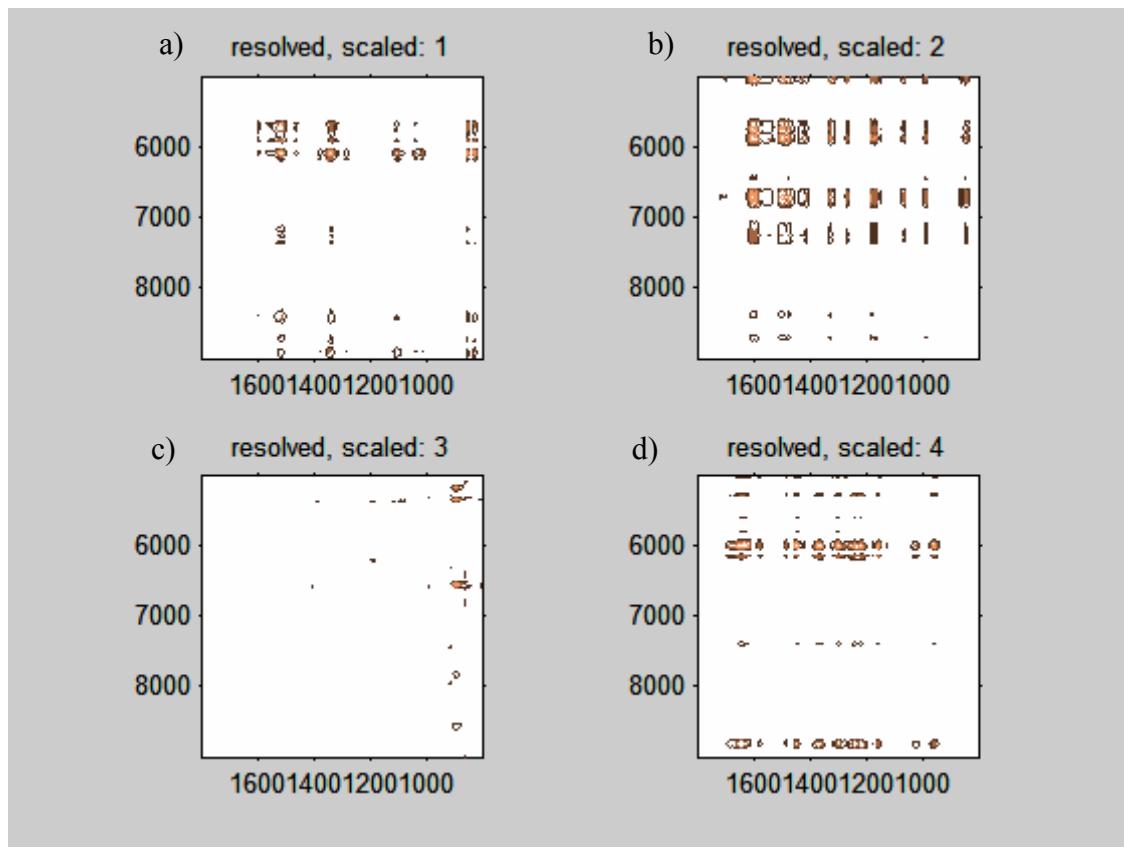


Figure 10. In (a) we see the results of predicting the mid-IR spectrum from the near-IR spectrum. In (b) the results are shown of predicting the near-IR spectrum from the mid-IR spectrum.

