PARAFAC questions: Spectroscopy workshop Granda May 2010.

Here is a list of questions brought up at the meeting with some short answers. Please also refer to the powerpoint silde.

Sheet 1
Problem: PARAFAC identifies my end members as outliers (e.g. torrential events)
Question: how to deal with these samples? Is there a problem with too heterogeneous data sets?

Two things to consider. Do these samples deviate as a result of high concentrations (same stuff but at higher levels) or are they very different in character in comparison to the other data?
Try
i) reducing the weight given to these samples by either (quick and dirty) dividing their fluorescence signal by a factor so that the fluorescence intensities are in a similar range to the other data. A more appropriate approach would be to mean center the data (divide all EEMs by their average, remember to multiply out again). Re do some simple models, does this have an effect on the components? (shapes).
ii) Remove these samples and see if you get a different result for simple 3-4 component models

Question: What is the primary validation approach.
Answer: There is no approach that will do for all datasets. One has to make a decision based on the combined result of many different results.

Sheet 2.
Question: How accurate is it to use PARAFAC to compare DOM peaks to where known compounds fluorescence to determine functional groups and if it isn't possible to use PARAFAC to ID functional groups, then what makes it so useful?
Answer: The approach is very good for this. The weakness is more with regards to the complexity of DOM rather than the effectiveness of PARAFAC. When working with simple mixture of known constituents the spectra obtained are the pure spectra of the fluorophores present. With DOM however there may be some issues (pH effects, metals, solvent effects etc…), with regards to the local environment of the fluorophore group which may influence the spectra. However if you have a compound in mind why not run some tests in the lab and see how its fluorescence is influenced by these factors. Additionally you can try to spike some of your samples with this compound and see how its fluorescence changes as a result of being in the DOM matrix rather than in pure MilliQ water.

Sheet 3.
Question: When you cut the data to remove Rayleigh lines I get a message "measurement error"?
Answer: Send me a mail with a screen shot of MATLAB when you get this error message. I have not seen this before and need to see what you have submitted. If this is happening on the tutorial data set make sure that you have not over written the original data. Try replacing the data with a copy of the original *.mat file from the DOMFluor zip file you downloaded.

Question: Have you encountered any system incompatibility with office 7 and MATLAB 2008 (shrinking graphs)?
Answer: As far as I can see this is a MATLAB 2008 and 2009 problem. Has nothing to do with Windows as far as I am aware. I am trying to find the time to work on it. Have not tried with 2010 yet. To get around this problem use the Surf plots instead. These do not shrink for some reason.

Question: Is there a minimum/ maximum/optimum number of components to build the model?
Answer: No. Determining the appropriate number of components is a difficult process and needs to be carried out with care and following all the steps in the tutorial.

The models are fit by a least squares regression and no "building" is carried out. Try to avoid using the term build. It insinuates that each component is found one after another, when they are actually determined simultaneously.

Question: Can PARAFAC fit oscillating trends (e.g. weekly oscillations)?
Answer: Yes. PARAFAC models the fluorescence data you provide irrespective of origin. Indeed including some natural variability into the data set might even make it easier to characterise with PARAFAC.

Sheet 4.
Question: What is the role of quenching?
Answer: I am a little unsure of the question here. If quenching is occurring then I presume that this will influence the appearance or disappearance of a component. This means that we should be careful when interpreting the loss of a component when analysing our data. This could be due to removal or quenching. Certainly an area worth more study.

Question: Role of acquisition parameters (speed, slit, step)
Answer: Rule of thumb is to try to measure all your samples with the same general set up, with regards to scan speed, and slit widths. Although I have not tried modelling data measured with different set ups. Also a good idea to have the same spectral coverage (ranges).

Parameter to choose the right magic number of components.
Answer: There is no magic number and no one test. Try to use the suite of approaches suggested in the tutorial and then make the decision on the basis of these results.

Sheet 5.
Question: How do I deal with removed scatter area? (NaN or zeros)
Answer: Combination of both. See the tutorial.

Question: Diagnostics: what are the best ways to validate the chosen number of components? (Core consistency, split half…..)
Answer: See points discussed above.

Question: Minimum number of components required? Can replicate samples be considered as individual samples in the model?
Answer: Do not use replicates to increase the data set size. They do not introduce any new information and may give problems with the split half validation (You could easily end up with one replicate in one half and the other in the other, making the approach in valid).

Sheet 6.
Question: How do determine the appropriate number of component? (Core consistency, split half…..) (See slide 12). You have to make a decision based on the results of the outlier analysis, residual analysis, split half validation, random initialisation analysis combined. Which number of components gives you a robust result. So if I gave you a new data set from the sample area you would be able to find exactly the same components.
Residual analysis: How do we make this less subjective? I do not know. This is not finished work. Feel free to develop an approach
How is it best to compare components between published models? (peak locations or overlap spectra) I would compare the spectral shapes. Most authors I am sure will send you their results

Sheet 7.
Question: What is the ideal number of components to build a PARAFAC model? What is the minimum?
Answer: See points above.

Question: I have 400 EEMs from 3 different watersheds, including WWTP effluent and 100 from quenching experiments.. would you a) build a model with all EEMs, b) build separate models? I would initially model separately.
Answer: See if there is any overlap between the results. Try combining afterwards and see if you get the same results.

Question: How do you determine where to cut the EEMs?
Answer: To a certain extent, by trial and error. A balance of not cutting too much valuable information away, and minimising the influence of any residual scatter effects or noisy data. Find the results which gives you the most sensible and robust results.

Question: In your opinion is it better to build an original PARAFAC model or fit data to existing models? Always best to first analyse your own data. If the components you find overlap with the earlier model then you can argue that it is OK to fit the other model to the new data.
Only fit an existing model to your data if you are 100% sure that there are no instrumental biases and that the data fall within the same system as the original model.
Try to avoid using the term “build”. See point earlier.

Question: Peak picking vs. PARAFAC? Benefits and disadvantages?
Answer: See my points in the power point slide. The one does not eliminate the other. Depending on the question being addressed and the state of your data set there are a range of approaches that can be applied.