Doc BDS EEM PARAFAC

Thursday, February 18, 2016 11:40 AM

BDS - PARAFAC-EEM - *Paving the way for an automated PARAFAC analysis*

http://djargon.azurewebsites.net/pdf/Doc11_BDSEEMPARAFAC

This walkthrough assumes that all spectral corrections have been performed on the data (already implemented in BDS or performed on the Aqualog directly):

1. Spectral Indices
    1. Spectra (abs. and fluo.) collected on most instruments have been corrected using the FDOMcorr toolbox.
    2. Spectra (abs. and fluo.) generated on the Aqualog are already corrected and uploaded in the system, ready to model. à See Aqualog Data import script to prepare the data for spectral indices export and PARAFAC modelling. Francois will provide this.
    3. Abs. and fluo. spectral indices are generated and provided to the user as a .csv or .xls file. The SpectralIndices.m (Aqualog) and FDOMcorr.m (other instruments) functions generate these indices, but call for Excel to write the file. Need a workaround here to write to .csv if .xls files cannot be generated by the BDS system.
2. Scattering and outliers removal and data normalization
    1. Removing noisy data

        ```
         SubData=subdataset(mydata,[],mydata.Em>600,mydata.Ex<240);
        SubData=subdataset(SubData,[],SubData.Em>600,SubData.Ex>500);
        ```

        Note: We may want to allow the user to specify the Emission and excitation boundaries or not.
    2. Removing Rayleigh and Raman scatters

        ```
        Xs=smootheem(SubData,[15 15],[15 15],[18 18],[18 18],[0 0 0 0],[20],3500,'');
        ```

        1st order Rayleigh
        1st order Raman
        2nd order Rayleigh
        2nd order Raman

        Note: Here we may want to allow the user to specify the range of scatter to remove. Also, it would be good to generate a 2D contour plot of each sample so the user can make sure that all scatters have been correctly removed. à EEMview function
    3. Data normalization

        From Xs, we enforce normalization

        ```
        Xpre=normeem(Xs);
        ```
    4. Perform exploratory test to search for outliers. Models with 3-7 comp. will be computed.

        Note: Here we may want to allow user to specify Nb. of components

        ```
        TestOutliers=outliertest(Xpre,[1,1],3:9,'nonnegativity',[],'at once');
        ```

        1. Calculate the leverage of each samples on the different models. Already present as a diagnostic tool in drEEM.
        2. Define a threshold above which a sample would be a good candidate for removal
        3. If the same sample is found to be an outlier in each model, remove the sample from the dataset

            ```
            XinModel3:9=subdataset(Xs,[outlier 1 outlier 2 outlier 3],[],[]);
            ```

The idea would be to great a new Xin dataset per model. 7 datasets total.

4. *A similar approach could be used for faulty WL in both the excitation or emission.*
    1. *Here the “zap” function could be used to remove these WL.* For future implementation

3. Validation phase

   Note: The determination of a good model is somewhat subjective and open to discussion, but here at least a first draft of what I have in mind.

    1. Random Initialization
        1. Run 3 to 9 comp. models, and identify the best run for each model

           ```
           [LSmodel3,convg3,DSit3]=randinitanal(XinModel3,3,5,'nonnegativity',1e-8);
           [LSmodel4,convg4,DSit4]=randinitanal(XinModel4,4,5,'nonnegativity',1e-8);
           [LSmodel5,convg5,DSit5]=randinitanal(XinModel5,5,5,'nonnegativity',1e-8);
           ```

           and so on.
    2. Split Half Analysis
        1. Let’s be a bit more vigorous here since the end user will have control on data quality at this point, and create 6 splits as suggested by Murphy et al.

           ```
           Split_Model3=splitds(XinModel3,[],4,'alternating',{[1 2],[3 4],[1 3],[2 4],[1 4],[2 3]});
           Split_Model4=splitds(XinModel4,[],4,'alternating',{[1 2],[3 4],[1 3],[2 4],[1 4],[2 3]});
           ```
        2. Create 3-9 models in each split Step 1: Check if the validation test passes

           ```
           A1_Model3=splitanalysis(Split_Model3,3,'nonnegativity',[],[],'A1');
           ```
        3. Attempt to validate

           ```
           splitvalidation(A1_Model3,3,[1 2;3 4;5 6]);
           ```
        4. Step 2: If the validation doesn’t work, then run this:

           ```
           A1_Model3 =splitanalysis(Split_Model3,3,'nonnegativity',[5 5 5 5 5 5],[1e-8 1e-8 1e-8 1e-8 1e-8 1e-8],'A1');
           ```
        5. Attempt to validate

           ```
           splitvalidation(A1_Model3,3,[1 2;3 4;5 6]);
           ```
    3. Note to self: Send Dmitry info about validation objects: Turns out that I was wrong, the objects that I was referring too are created during the export, not validation. However, a few figures will be generated at this point showing the different splits and model loadings. These figures need to be printed out and forwarded to the user.

4. Model Export

   Note: Here we will always use the projected function. This means that when there are outliers present in the dataset, the model will be retrofitted to those outliers. If no outlier present, then the output will be similar in both the modelled data per se and the projected data. So, no harm here.

   In the example below, a five component model (LSmodel5) is being projected to the full dataset where the scatter has been removed, but outliers still present. This should always be Xs.

    1.
       ```
       [F5,EmSpectra,ExSpectra,Ff,P5]=modelout(LSmodel5,5,'Nameofthedataset.xls',Xs);
       ```
    2. Objects created are F5, EmSpectra, ExSpectra, Ff, P5, and an excel file that needs to be forwarded to the user. Here again, .xls to .csv workaround needed

   FG: I did some cleaning in the list below for the FDOM indices as many of these are either outdated or duplicates of the same thing.

   Note: some of the FDOM peaks are often quite broad, and for now I used either the average position or my own experience as the coordinates of each peak. In a future release, we can add a function that finds the max peak in both excitation and emission, but that will require some programming and licensing issue again. Feasible, but not a priority for now since the whole idea of PARAFAC is to identify these peaks (and others).

Index = ratios of fluorescence excitation:emission pairs from Chapter 9 of Aquatic Organic Matter Fluorescence. Table 9.1

MAXs are for typical peaks listed in Table 3.1 from same book

HIX_{EM}	254/435-480:254/300-345
BIX (new freshness index)	310/380:max310/420-435
FI_{NEW}	370/470:370/520
Tyrosine-b (peak B)	275/305
Tryptophan-b (peak T)	275/340
N (peak N)	280/370
M (peak M)	320/410
Cb (peak C)	350/450
C+a (peak A)	250/440