EXCITATION-EMISSION-MATRIX'S CONVERSION, CORRECTION, COGNITION, COMPARISON, AND CALCULATION (EFC) V1.0

MAIN FUNCTIONS AND THEIR CALCULATION METHODS

Dr. Wei He[¢,¶]

Prof. Fuliu Xu[¶]

[¢]School of Water Resources and Environment, China University of Geosciences (Beijing)

[¶]College of Urban and Environmental Science, Peking University

CONTENTS

The main functions of efc software can implement pretreatment, preliminary analysis, and deep analysis of dataset. The functions of pretreatment mainly include blank subtraction of EEM, correction of the inner filter effect, the Raman normalization, calibration with quinoline sulfate, and elimination of Raman & Rayleigh scattering. The preliminary analysis mainly includes calculation of absorption spectrum index and fluorescence spectrum index. The deep analysis of data mainly includes preliminary PARAFAC analysis, PARAFAC analysis without outliers, model optimization and validation, and model comparison and fitting.

1. Pretreatment of data

1.1. E Pretreatment of EEM using interpolation

The emission wavelengths of EEM measured by Aqualog (HORIBA) don't have equal interval and are not integers. To make those data comparable with data measured by other instrument, which produce data with equal interval, and to acquire fluorescence at specific excitation and emission wavelength, the interpolation is needed. Hence, we write a self-built code interp_EEM based on function interp2 in MATLAB toolboxes. The linear method is employed in the function interp_EEM. Due to the limitation of interp2, the interpolated values could not exceed the original excitation and emission wavelength ranges, otherwise the interpolated values might be assigned as NaN, it will not be calculated later. To interpolate the emission spectrum at specific excitation, the function interp is employed. The interpolated value's range could not exceed the original.

1.2. Instrument spectral correction

The excitation emission correction factors are typically employed to correct the measured EEM from various instrument to make them comparable with each other. Murphy et al. (2010) investigated the excitation emission correction factors of popular fluorescence spectrometers made by Varian, Horiba Jobin Yvon, Hitachi, Photon Technologies International, and PerkinElmer and found the instrument spectral correction factors of Varian and Horiba varied with the instruments and spectral correction factors are similar for the instruments manufactured by the sample company. The following data handling refers to the MATLAB code FDOMcorrect.m writtern by by Murphy et al. (2010). The protocol of the EEM correction code is shown in Figure 1.

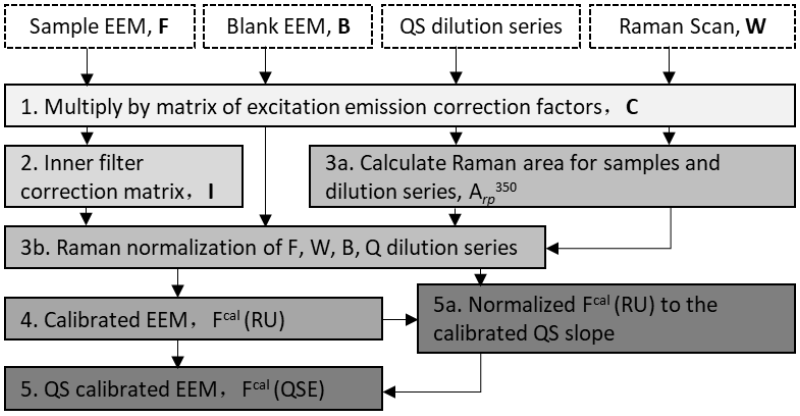

Figure 1 The EEM correction protocol proposed by Murphy et al. (2010)

The measured EEMs such as sample (**F**) and blank (**B**) were corrected though multiplication of them by the EEM-sized, instrument-specific correction matrix (**C**). The matrix **C** was calculated the outer product of the vector of excitation correction factors, $\mathbf{f}_{\lambda ex}^{Excorr}$, and the transposed vector of emission correction factors, $\mathbf{f}_{\lambda em}^{Emcorr}$, (Eqs. (1) – (3)). The correction factor of the water Raman scan spectrum $\mathbf{W}_{350,\lambda em}$ is the emission correction factors corresponding λ_{ex} = 350 nm of **C** and the correction factors of slope $m_{350/450}$ of QS dilution series is the correction factor at $\lambda_{ex}/\lambda_{ex}$ = 350/450 nm of **C**. Their correction refers to Eqs. (4) and (5).

$$\mathbf{C} = \mathbf{f}_{\lambda ex}^{Excorr} \times (\mathbf{f}_{\lambda em}^{Emcorr})^{\mathrm{T}} \quad (1)$$

$$\mathbf{F}^{corr}(AU) = \mathbf{C} \cdot \mathbf{F} \quad (2)$$

$$\mathbf{B}^{corr}(AU) = \mathbf{C} \cdot \mathbf{F} \quad (3)$$

$$\mathbf{W}_{350,\lambda em}^{corr} = \mathbf{C}_{350,\lambda em} \cdot \mathbf{W}_{350,\lambda em} \quad (4)$$

$$m_{350/450}^{corr} = \mathbf{C}_{350/450} \cdot m_{350/450} \quad (5)$$

1.3. Inner filter effect correction

The inner filter effect correction follows the proposed models by Parker and Barnes (1957). And the correction matrix **I** is calculated by Eq (6).

$$\mathrm{I}_{\lambda_{ex},\lambda_{em}} = 10^{(L_{ex}A_{\lambda_{ex}} + L_{em}A_{\lambda_{em}})/2} \quad (6)$$

where absorbance $A_{\lambda ex}$ and $A_{\lambda em}$ are acquired using absorption spectrophotometry, and their wavelength ranges correspond with the excitation and emission wavelength ranges of EEM. L_{ex} and L_{ex} are the light path (cm) of the cuvette in the direction of excitation and emission. When the cross section of the cuvette is a rectangle with unequal length and width, L_{ex} and L_{ex} are different. When the cross section of the cuvette

is a square and its width is 1 cm, the length and width are both 0.5 cm. Therefore, the Eq. (6) can be simplified as the Eq. (7). Inner filtering is accounted for by element-wise multiplication of each spectrally corrected EEM, $\mathbf{F}^{corr}$, by correction matrix **I** as shown in Eq. (8).

$$I_{\lambda_{ex},\lambda_{em}} = 10^{(A_{\lambda_{ex}} + A_{\lambda_{em}})/2} \quad (7)$$

$$\mathbf{F}^{IFE} = \mathbf{I} \cdot \mathbf{F}^{corr} \quad (8)$$

1.4. Raman normalization

Raman normalization is calculated through dividing the fluorescence intensity data by Raman peak area, which make each EEM data comparable. The sample EEMs, blanks, and dilution series slope are required to be Raman normalized as shown in Eqs. (9) – (11), which will produced corrected fluorescence in Raman Unit.

$$\mathbf{F}^{cal}(RU) = \mathbf{F}^{IFE} / A_{rp}^{350} \quad (9)$$

$$\mathbf{B}^{cal}(RU) = \mathbf{B}^{corr} / A_{rp}^{350} \quad (10)$$

$$m_{350/450}^{cal}(RUQSE^{-1}) = m_{350/450}^{corr} / A_{rp}^{350} \quad (11)$$

where the corrected Raman peak area (A_{rp}^{350}) is a scalar value specific to each sample and calculated in Eq. (12).

$$A_{rp}^{350} = \int_{381}^{426} W_{350,\lambda_{em}}^{corr} d\lambda_{em} \quad (12)$$

where A_{rp}^{350} is the baseline-corrected area under the curve $\mathbf{W}_{350,\lambda em}^{corr}$ that is bounded by the emission wavelengths 381 and 426 nm. The emission wavelength range might varied among samples and blanks because of the fluctuating PMT signal strength and is required to adjust according to the practical situation.

1.5. Blank subtraction

In the case of a larger background, the blank subtraction as shown in Eq. (13) is necessary, but for the dilution sample, that subtraction remains doubtful because it might not improve fluorescence measurement accuracy (Zepp et al., 2004).

$$\mathbf{F}^{cal}(RU) = \mathbf{F}^{cal}(RU) - \mathbf{B}^{cal}(RU) \quad (13)$$

1.6. Calibration with quinoline sulfate

Generally, the Raman normalized and spectrally corrected EEMs ($\mathbf{F}^{cal}$) are calibrated against the QS dilution series (0 – 100 ppb), producing EEMs in QSE (1 QSE = 1 ppb quinoline sulfate in 0.1 N sulfuric acid, measured at $\lambda_{ex}/\lambda_{ex}$ = 350/450 nm). The

Raman calibrated EEMs are divided by the calibrated slope (Eq. (11)) as shown in Eq. (14). Since the concentrations of QS dilution series are so long that the weak inner filter effect can be neglected, the inner filter correction is not conducted for QS dilution. The blank subtraction only affects the intercept of standard curve but does not affect the slope, so the blank subtraction of QS dilution series is not considered (Murphy et al, 2010)。

$$\mathbf{F}^{cal}(QSE) = \mathbf{F}^{cal}(RU) / m_{350/450}{}^{cal}(RUQSE^{-1}) \quad (14)$$

1.7. Elimination of Raman & Rayleigh scattering peaks

The scattering peaks in the EEM of NOM mainly include the first Raman scattering, the first Rayleigh scattering, the second Rayleigh scattering, and sometimes the second Raman scattering (Figure 2). Zepp et al. (2004) wrote MATLAB code to eliminate those scattering peaks. Their basic idea is (1) to assign the data in the scattering peak region as NaN and (2) to interpolate those NaN regions using MATLAB functions interp1 and interp2. In DOMfluor, developed by Stedmon and Bro (2008), values in the scattering peak area are given to NaN or 0, so those data do not participate in PARAFAC analysis. The former method makes the data more complete, but the defect is that the interpolation changes the data greatly. And the interpolated data can hardly reflect the real situation. Although the data in the latter elimination method is incomplete, the data used is the real data, which is more rigorous. Both methods are acceptable in subsequent EEM analysis, therefore the software includes both two ways of eliminating scattering peaks.

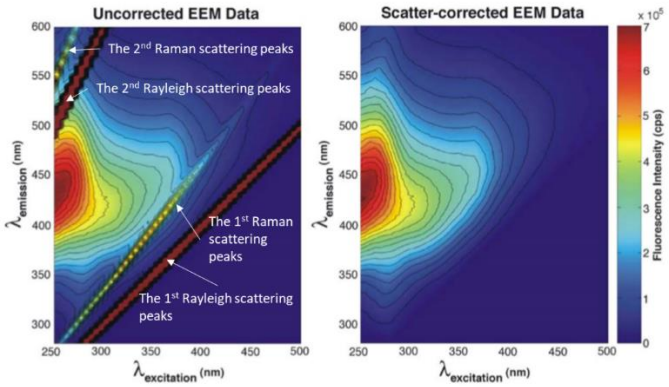

Figure 2 Elimination of Raman and Rayleigh scattering by interpolation (Zepp et al., 2004)

2. Preliminary analysis of data——calculation of indicators

2.1. Absorption spectrum indicators ---- Absorptivities and their ratios

NOM is a group of organic chemicals containing abundant conjugate structures.

Therefore, its aqueous solution has a strong signal value in the ultraviolet region of the absorption spectrum. The absorbance coefficient (absorptivity) (e.g., a_{254}, a_{260}, a_{360}, a_{360}) is a typical indicator that characterize NOM's conjugate structures, and the calculation method is shown Eq. (15).

$$a_\lambda\ [m^{-1}]=2.303 \cdot Abs_\lambda / L \quad (15)$$

where, L is the optical path of the cuvette (usually 0.01 m).

Absorptivity ratio E2/E3, E2/E4, E4/E6 are ratio of a_{254} to a_{365}, ratio of a_{254} to a_{436}, and ratio of a_{465} to a_{665}, respectively. They can indicate the humification degree of NOM, NOM sources, and the condensation and conjugate degree of aromatic carbon skeleton, respectively (Ytow et al., 1996; Jaffe et al., 2004).

2.2. Absorption spectrum indicators ---- Spectral slopes and their ratios

Spectral slope, $S_{\lambda 1-\lambda 2}$, is calculated by fitting the absorption spectrum with the nonlinear regression equation (16). The parameter S is (Battin, 1998), if the λ unit is μm, the unit of S will be μm^{-1}.

$$a_\lambda = a_{\lambda 0} e^{-S(\lambda-\lambda 0)} \quad (16)$$

$$\ln a_\lambda = S(\lambda-\lambda_0) + \ln a_0 \quad (17)$$

where λ_0 is the reference wavelength, i.e., λ_1; λ has ranges of $\lambda_1 < \lambda \leq \lambda_2$. When conducting fitting calculation, we transformed the Eq. (16) by taking the logarithm of both sides of equation to Eq. (17). Then the linear formula Eq. (17) is used to fit the spectrum data.

The ratio of spectral slopes $S_{275-295}$ to $S_{350-400}$ is the derivative indicator S_R. Its lower value indicates the input of solubility organic matters with high molecular weight, strong aromatic, and sources of vascular plants (Spencer et al., 2012).

2.3. Absorption spectrum indicators ---- Spectral areas and their ratios

The areas under the absorption spectrum of 260 ~ 280 nm, 460 ~ 480 nm and 600 ~ 670 nm are integrated and defined as spectral area Q_2, Q_4 and Q_6. Their ratios are Q_2/Q_4, Q_2/Q_6, and Q_4/Q_6, respectively. The Spectral area is similar to absorbance coefficient, but covers more chemical information rather than a single wavelength. The spectral area ratios can indicate the discrimination of lignin, the humification degree, condensation of aromatic compounds (He et al., 2016; Albrecht et al., 2011. The calculation of spectral area refers to Eq. (18).

$$Q_{\lambda_1-\lambda_2} = \int_{\lambda_1}^{\lambda_2} Abs(\lambda)d\lambda \quad (18)$$

where λ_1 and λ_2 are wavelength boundary of the absorption spectrum.

2.4. Fluorescence spectrum indicators ----Fluorescent peaks

The conventional fluorescent peaks are found in fixed excitation emission wavelength regions or at a fixed excitation and emission wavelength. Coble (1996) proposed the excitation (Ex) and emission (Em) wavelength ranges for each fluorescent compounds. The tyrosine-like (Peak B) has Ex ranges of 250 ~ 300 nm and Em ranges of 280 ~ 330 nm; the tryptophan-like (Peak T) has Ex ranges of 250 ~ 300 nm and Em ranges of 330 ~ 380 nm; the low-excitation-wavelength (UVC) humic-like (Peak A) has Ex ranges of 250 ~ 300 nm and Em ranges of 380 ~ 480 nm; the marine humic-like (Peak M) has Ex ranges of 300 ~ 320 nm and Em ranges of 380 ~ 420 nm. In this software, a new fluorescent peak scope (ACSTDB) is proposed under the summary of the studies in recent two decades (He et al., 2016). The ACSTDB includes fulvic-like (Peak A) with Ex ranges of 200 ~ 275 nm and Em ranges of 380 ~ 550 nm, humic-like (Peak C) with Ex ranges of 380 ~ 550 nm and Em ranges of 380 ~ 550 nm, low-excitation-wavelength tryptophan-like (Peak S) with Ex ranges of 200 ~ 250 nm and Em ranges of 330 ~ 380 nm, high-excitation-wavelength tryptophan-like (Peak T) with Ex ranges of 250 ~ 300 nm and Em ranges of 330 ~ 380 nm, low-excitation-wavelength tyrosine-like (Peak D) with Ex ranges of 200 ~ 250 nm and Em ranges of 280 ~ 330 nm, and high-excitation-wavelength tyrosine-like (Peak B) with Ex ranges of 200 ~ 250 nm and Em ranges of 280 ~ 330 nm. The protocol of peak finding is (1) to smooth the noise peaks in the EEM and (2) to find the maximum peaks in both excitation and emission dimension. The software only gives a single peak values, and the potential other lower peaks are not consider in this version.

2.5. Fluorescence spectrum indicators ---Fluorescent volume

The fluorescent volume, also known as fluorescence regional integration, is the volume beneath the EEM surface in a specific excitation-emission wavelength region. It is calculated using Eq. (19) proposed by Chen et al. (2003). For discrete data, the volume was expressed by Eq. (20).

$$\Phi_i = \int_{ex}\int_{em} F(\lambda_{ex},\lambda_{em})d\lambda_{ex}d\lambda_{em} \quad (19)$$

$$\Phi_i = \sum_{ex}\sum_{em} F(\lambda_{ex},\lambda_{em})\Delta\lambda_{ex}\Delta\lambda_{em} \quad (20)$$

where i is a specific excitation emission wavelength matrix region, which is divided according the fluorescent peaks' region, $F(\lambda_{ex}, \lambda_{em})$ is the fluorescence intensity

at a specific excitation-emission wavelength pair, $\Delta\lambda_{ex}$ and $\Delta\lambda_{em}$ are wavelength intervals, which depend on the measured or interpolated settings.

Normalized excitation-emission area volumes ($\Phi_{i,n}$) and percent fluorescence response ($P_{i,n}$) are calculated according to Eq. (21) and Eq. (22), respectively.

$$\Phi_{i,\mathrm{n}} = MF_i \Phi_i \tag{21}$$

$$P_{i,\mathrm{n}} = \Phi_{i,\mathrm{n}} \sum_{i=1}^{m} \Phi_{i,\mathrm{n}} \tag{22}$$

where MF_i is calculated through dividing sum of all the projected excitation-emission area by projected excitation-emission area of region "*i*".

2.6. Fluorescence spectrum indicators ----Fluorescent index

The fluorescent indexes calculated in the software includes fluorescence index (FI) fluorescence indices (McKnight et al., 2001), modified fluorescence index (FI2) (Jaffe et al., 2008), the humification index (HIX) (Zsolnay et al., 1999), biological index (BIX) (Huguet et al., 2009) and normalized humification index (HIXc) (Ohno, 2002). The indexes' calculation refer to Eqs. (23) – (27).

$$FI = F_{370/450} / F_{370/500} \tag{23}$$

$$FI_2 = F_{370/470} / F_{370/520} \tag{24}$$

$$HIX = \sum F_{254/435 \to 254/480} / \sum F_{254/300 \to 254/345} \tag{25}$$

$$HIX_c = \sum F_{254/435 \to 254/480} / (\sum F_{254/300 \to 254/345} + \sum F_{254/435 \to 254/480}) \tag{26}$$

$$BIX = F_{310/380} / F_{310/430} \tag{27}$$

2.7. Fluorescence spectrum indicators ----Specific fluorescence

The software can also give the fluorescence intensity at a specific excitation-emission wavelength pair, which might be used to calculated other self-defined fluorescent index or serve as a bio-marker. It is a potential expansion function. It should be noted that the specific fluorescence might not be measured in the original data. Therefore, the software employs the two-dimensional interpolation method to obtain that value. The previous fluorescence, which is not measured, is also obtained by that way.

3. Deep analysis of data——PARAFAC analysis

3.1. Preliminary PARAFAC analysis

At present, the core code of PARAFAC analysis employ the open source MATLAB toolbox n-way developed by Bro (1998), which is described in detail in Figure 7. The function of preliminary PARAFAC analysis is as follows:

```
[Model iter err Corcondia]=parafac(EEM.X,NumFac,...
       [str2double(get(handles.edit15,'String')) 1 0 1 NaN...
       str2double(get(handles.edit16,'String'))],...
       [constrict constrict constrict]);
```

In the parafac function (Figure 3), the convergence criterion, maximum number of iteration, and constraint value are obtained through the efc software interface. Before the preliminary PARAFAC analysis, the EEM can be cut or eliminated using the scattering elimination function mentioned above.

```
[Factors,it,err,corcondia]=parafac(X,Fac,Options,const,OldLoad,FixMode,Weights)
```

INPUT	
x	Input array, three- to N-way, and three-way for EEM, i.e., samples*excitation*emission 3D data.
Fac	Number of factors/components sought.
Options	Optional parameters, six parameters, If not given or set to zero or [], defaults will be used. Options(1) ---Convergence criterion, default value is 1e-6, but difficult data might require a lower value. The value can be set in efc software. Options(2) ---Initialization method, default value is 0, fit using DTLD/GRAM for initialization; 1=fit using SVD vectors for initialization; 2=fit using random orthogonalized values for initialization; 10=fit using the best-fitting models of several models. The value cannot be set in efc software. 2 is assigned when conduct PARAFAC analysis with random initialization; 1 is assigned in other PARAFAC analysis. Options(3) ---Plotting options. 0 =no plots; 1 or 2 = produces several graphical outputs; 3= as 2 but no core consistency check. The value cannot be set in efc software. 0 is assigned for calculation efficiency. Options(4) ---Scaling, default value is 0 or 1, scaling; 2 = no scaling applied. The value cannot be set in efc software. 0 is assigned for calculation efficiency. Options(5) ---How often to show fit, default value is 10, NaN = no outputs are shown. The value cannot be set in efc software. NaN is assigned for calculation efficiency. Options(6) ---Maximal number of iterations, default value is 0, indicating integrations is 2500. The value can be set in efc software, the default value is 2500.
const	Constraints. 0 => no constraint; 1 => orthogonality; 2 => non-negativity; 3 => unimodality (and non-negativitiy). The value can be set in efc software, the default value is 2, indicating non-negativity.
OldLoad	Initial guess of the loadings. The parameter is assigned when fitting a known PARAFAC model, the three loadings is obtained using fac2let. The value can be set in Model fitting module of efc software.
FixMode	Score settings. For a three-way model, [x1 x2 x3] find the scores of a data set given the loadings, xi = not require old values hence estimated, xi = 1 requires old values given. The value cannot be set in Model fitting module of efc software, the assigned value is [0 1]
Weights	Weighted regression. If a matrix of the same size as X is given, weighted regression, is performed using the weights in the matrix Weights. The value is not assigned in efc software
OUTPUT	
Factors	PARAFAC estimate of loadings in one matrix. The important results of PARAFAC analysis.
it	Number of iterations used. Can be helpful for checking if the algorithm has converged.
err	The fit of the model = the sum of squares of errors (not including missing values).
Corcondia	Core consistency test. Should ideally be 100%. If significantly below 100% the model is not valid.

Figure 3 The input and output of parafac function in the n-way toolbox and their settings in efc software

3.2. PARAFAC analysis without outliers

After the preliminary PARAFAC analysis, we can plot loading and leverage of different models. The smooth degree of loading curves of excitation wavelength and emission wavelength can indicate the quality of the PARAFAC models to some extent. Generally, the smooth and non-negative loading plot declare good PARAFAC model. The leverage plot of samples indicates which sample might be a potential outlier. Those outliers might be caused by questionable measurement. Therefore, it is suggested to plot the EEM contour graph and check its completeness. If there was a problem for that EEM, it is required to be measure again or deleted when conducting PARAFAC analysis. The PARAFAC analysis without outliers employed the MATLAB code as follows:

```
EEM=deletEEM(EEM,Str2WL(get(handles.edit19,'String')),0);
[Model iter err Corcondia]=parafac(EEM.X,NumFac,...
        [str2double(get(handles.edit15,'String')) 1 0 1 NaN...
        str2double(get(handles.edit16,'String'))],...
        [constrict constrict constrict]);
```

where the deletEEM function can realize removing the sample when its sequence number is assigned.

3.3. Model optimization and validation----Sum of the squared errors, SSE

After PARAFAC analysis, we employed several ways to select the best validated model and results with high accuracy in the present and following sections. The PARAFAC analysis above generally includes many models with various components (> 2). Then the model optimization and validation will be performed. This section use the statistic indicators, sum of the squared errors (SSE) between observed value and simulated value. The SSEs in three dimension, samples, excitation wavelength, and emission wavelength, are calculated to investigate which model has smaller SSE. Generally, PARAFAC model with larger number of components will has smaller SSEs. However, SSE differences between models will become smaller and smaller with the increase of the component number. In this case, it is necessary to further select the best model through ways of core consistency and split half validation.

3.4. Model optimization and validation----Core consistency

As shown in Figure 7, the output of each PARAFAC analysis includes SSE, err, the final iteration number, iter, the core consistency diagnostic value, Corcondia, and the derivation parameter, explained variations. Among them, core consistency is the most effective parameter to select the best models (Bro and Kiers, 2003). Generally, core consistency diagnostic value of PARAFAC model with more components will be

far away from 100%. Nevertheless, the model is much closer to the measured EEM when the diagnostic value is approaching to 100%. If the diagnostic value is significantly lower than 100%, the model might be valid. Sometime diagnostic value is too small to decide which model is better. Thus, it is necessary to combine with SSE, explained variations, and even number of iterations. The core consistency indicates that the model with less component are much better based on diagnostic value within a certain scope of models. It is different from the SSE, which indicates that the model with more component is better. Therefore, it is reasonable to combine those two parameters to select the best model in practice (He et al., 2016).

3.5. Model optimization and validation----Split half validation

The original data will be split (divided) into four groups, which will be randomly assembled in pairs, i.e., four halves according to Figure 4. In the 1st split half group, Half 1 is the calculation data and Half 2 is the verification data; in the 2nd split half group, Half 3 is the calculation data and Half 4 is the verification data. The consistency between calculation data and verification data will be tested by Turky consistency diagnostic value. Strictly, for both two split half groups, the PARAFAC model of verification data should be consistent with that of calculation data. Then we declare that that model has passed the split half validation (Stedmon and Bro, 2008).

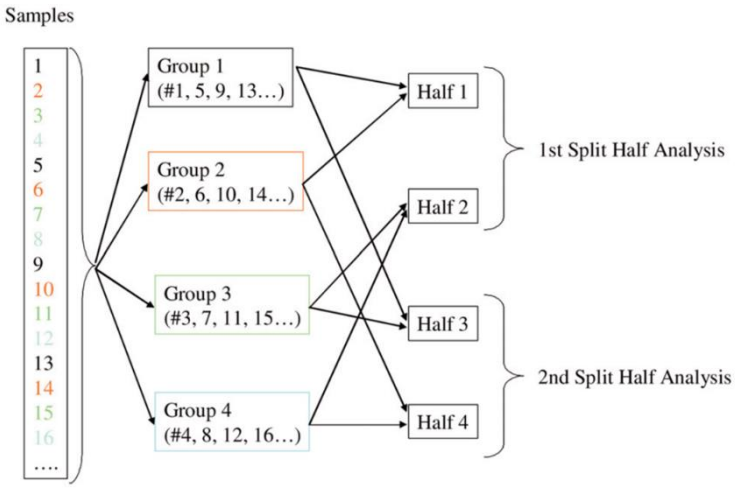

Figure 4 Data split handling before split half validation

3.6. Model optimization and validation----PARAFAC analysis with random initialization

In order to maintain the consistency of each PARAFAC analysis above, the PARAFAC employs singular value decomposition (SVD) vectors for initialization.

After confirming the best model, the model need to be optimized by employing random orthogonalized values for initialization to improve the fit of the model. The software realizing the PARAFAC analysis with random initialization through the function proposed by Stedmon and Bro (2008) as follows:

```
[Model iter err Corcondia]=parafac(EEM.X,NumFac,...
               [str2double(get(handles.edit15,'String')) 2 0 1 NaN...
               str2double(get(handles.edit16,'String'))],...
               [constrict constrict constrict]);
```

After 10 runs of PARAFAC analysis with random initialization, the result with the least err value will be the finally optimized and validated PARAFAC models.

3.7. Model comparison

After getting the PARAFAC model, the researchers can compare their PARAFAC model with other models in literatures. Most researches carried out the model comparison by the peaks' wavelengths of components' EEMs. Parr et al. (2014) employed mTTC to compare for the first time, which it possible to compare user model with the library models. The software can realized the PARAFAC model comparison between target model and the library models and between target model and the imported model by users based on the MATLAB code written by He and Hur (2015). The model comparison functions include comPARAFAC and comPARAFAC2. The former can compare the target model with the library models, and the latter can compare the target model with imported model by users. The model library in this software also gives explanations of the various components of different models. In addition the library can also be extended by the EEM-PARAFAC module of efc software.

3.8. Model fitting

Some researchers are particularly interested in others' PARAFAC models, hoping that their EEM data could be fitted using those models. Cory and McKnight (2005) constructed a 13-component PARAFAC model (CM model) using a big EEM dataset. Some components were similar to the known substances such as semiquinone and oxidized quinone. Therefore, many researchers expect to use CM model to fit their own data, and get the PARAFAC scores of each component. Fellman et al. (2009) employed two models including CM model to fit their EEM data. After getting the scores of each component, the redox index (RI) was calculated using the scores of some components. The RI was further used to judge the two models. The model fitting functions in the present software include fitPARAFAC and fitPARAFAC2. The former can fit EEM dataset using the library models, and the latter can fit EEM dataset using imported model by users. The core code of model fitting is as follows:

```
[Model iter err Corcondia]=parafac(EEM.X,DataBase{nd,4},...
       [p1 1 0 NaN p2],[p3 p3 p3],OldLoad,[0 1 1]);
```

3.9. Overall quality of data fitting ---- Relative SSE (RSSE)

Relative sum of square error (RSSE) are used to evaluate the differences between the simulated EEM by PARAFAC model ($X_{est,i}$) and the measured EEM ($X_{obs,i}$). Generally, the smaller RSSE indicates the better overall data fitting.

$$RSSE = SSE / SSX = \sum_{i=1}^{n} (X_{obs,i} - X_{est,i})^2 / \sum_{i=1}^{n} X_{obs,i}^{\ 2} \qquad (27)$$