Supporting Information Accompanying

"Estuarial Fingerprinting through Multidimensional Fluorescence and Multivariate Analysis"

Gregory J. Hall[†]

Kerin E. Clow[‡]

Jonathan E. Kenny[*]

This supporting information contains two sections. The first is a description of methods used, and relevant data to the validation of our NPLS-DA models. The second is the description and relevant data to our method of detecting outliers in our PARAFAC model.

Supporting Information

Estuarial Fingerprinting through Multidimensional Fluorescence and Multivariate Analysis

Gregory J. Hall, Kerin E. Clow, and Jonathan E. Kenny

NPLS-DA model validation

During the calculation of an NPLS model, the number of Latent Variables (LVs) must be selected. This is accomplished through cross validation. Multiway cross validation (CV) is performed on the model to determine the residual errors in both predictive ability (F) and the model's fit to the data (E). The multiway CV method performs iterative calculations to provide the Root-Mean-Squared Error of Cross Validation (RMSECV) and Root-Mean-Squared Error of Calibration (RMSEC). The appropriate number of LVs is chosen to provide maximum predictive power before the model begins to overfit the data. The number of LVs is determined by a balance between a minimal RMSECV value and a significant decrease from the previous LV in the RMSEC for the independent variables. The success of a model is also indicated in the percent variance captured in X and Y blocks. A significant improvement in the variance captured in either variable block should occur with each consecutive LV. The NPLS-DA regression models utilize the multivariate nature of EEM fluorescence response of very similar but unique sets of samples to discriminate and classify them most efficiently for calibration and prediction of new data.

A single value is given for the error in the predictive ability of the entire model, which is determined by class separation and misclassification probability of test data. The RMSECV value increases as sites of interest are added to the model. For models of the selected number of LVs, a separate RMSEC value is provided as a measure of the error in the model's fit to the data from each sample site (Table S1). The spectral differences between sites L and R are small and the model has some difficulty fitting the data. In Model LRHC, the differences between H and C are comparably smaller than between any other sites and the RMSEC value increases slightly for these sites.

Table S1. Error of Predictive Power (RMSECV) and Error of Fit to Data (RMSEC) for Each Model

Model	RMSECV	RMSEC (L)	RMSEC (R)	RMSEC (H)	RMSEC (C)
LRHC (6 LV)	0.708	0.162	0.190	0.224	0.232
LC (5 LV)	0.281	0.0904	--	--	0.0904

-- = not calculated.

The success of the calibration model is determined by these statistics resulting from multiway cross-validation using the leave-one-out method. The number of latent variables (LVs) for each model was chosen to minimize the cumulative Predictive Residual Sum of Squares (PRESS) result (Table S2). In each case, PRESS values reach a minimum corresponding to the number of latent variables that minimize the RMSECV and provide a significant decrease in the RMSEC as described above. Model LC was most successful based on five latent variables, while Model LRHC was based on six. The error increases with additional sites in the calibration model, as expected with limited geographic separation and small variations between sample sites.

Table S2. Cumulative PRESS Results for Each Model

Model	LV 1	LV 2	LV 3	LV 4	LV 5	LV 6	LV 7	LV 8	LV 9
LRHC	19.2	19.0	18.6	17.9	16.9	**16.0**	17.2	18.0	20.0
LC	1.03	1.53	1.25	1.39	**1.27**	1.92	3.36	--	--

The amount of co-variance captured by the calibration models can be calculated as a final check of validity. Values of over 90% variance explained in both X and Y blocks defines a well-fitted calibration model (Table S3). Model LC, with greater than 91% variation captured in both X and Y blocks, shows that the calibration step for sites L and C, separated by only six miles, was successful and provides support for this sample identification technique using NPLS-DA. As indicated by the dramatic decrease in percent variation captured in the Y block of Model LRHC, the calibration of site variation was not as rigorous and the predictive ability for a model containing all four sites, including sites H and C, which are within one mile of each other, becomes limited. The waters from L, C and R all mix into the Boston Harbor, which produces more predicted Y values in the false negative range and increases the difficulty of correctly classifying samples from H. The percent variation captured in the X block remains high for both models.

This result signifies that the model is able to identify latent structures in the fluorescence data to define most of the variance between the sites of interest using the calibration samples provided.

Table S3. Percent Variation Captured in Each Model

Sites of Interest	Number of LVs	Percent Variation Captured X-Block	Percent Variation Captured Y-Block
L,C	5	96.50	91.08
L,R,H,C	6	96.38	65.76

PARAFAC

Factor number selection

The number of factors to include in a PARAFAC model can be determined in several ways. In some instances additional factors are nearly identical to others and are unnecessary. Recently Bro et al. have devised a method where the core consistency of the PARAFAC model is used as a measure of factors to utilize.(*1*) In this method the number of factors that comes closest to a 100% consistent core should be used in the final model. All models in this work were fit for two, three or four components. The models with the highest number of components with a core consistency of over 97% were kept. All models found core consistencies of over 97% for 3 factors. One location, C, had a 97% consistency for a four component model; however, two factors in that model were practically identical, so a three factor model was kept for interpretation.

Outlier Removal

Outliers were identified by using RIP and IMP plots as described by Riu et al.(*2*), where each EEM represents the data on a particular date in a specific location. RIPs identify EEMs that differ greatly from the other EEMs in the data set.(*2*) They can be interpreted to have large differences in their loadings in the spectroscopic axes. IMPs identify EEMs that have vastly different scores in the date axis and this analysis is done for each factor.(*2*) While there is not space to show every RIP and IMP plot, the case of Mystic River, having one outlier, is shown below (Figures S1-S2).

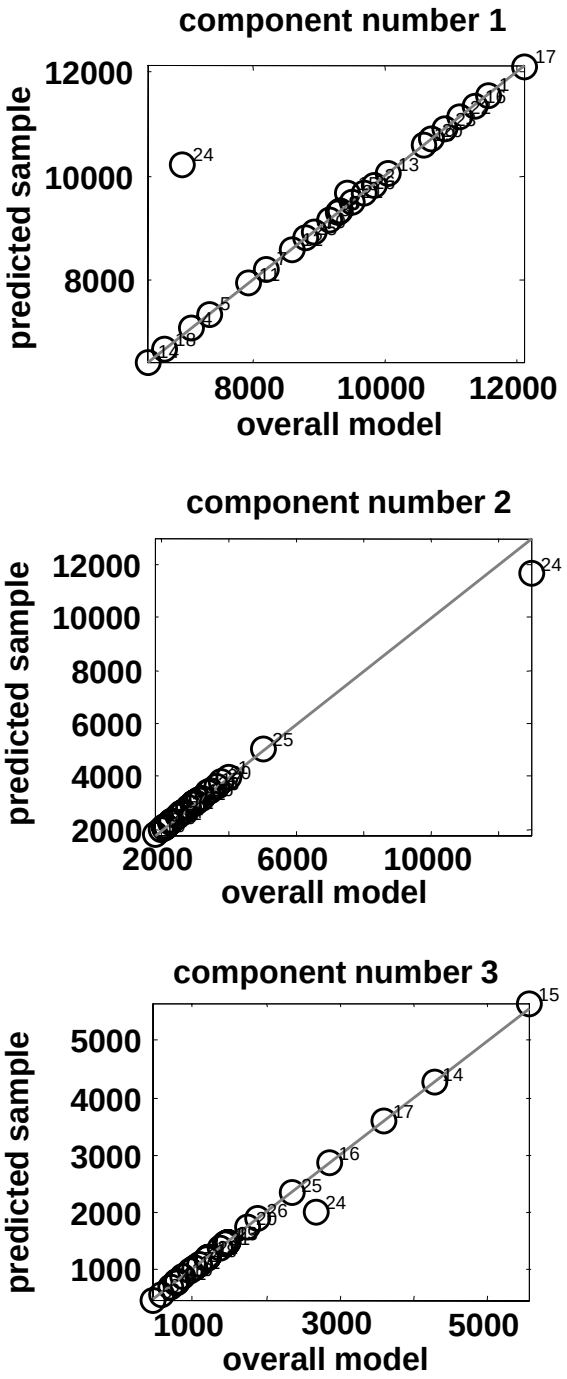

Figure S1 – Identity Match Plots (IMP) for Mystic River. Sample 24 is a significant distance from the identity line in all three components (factors), and therefore can be considered an outlier.

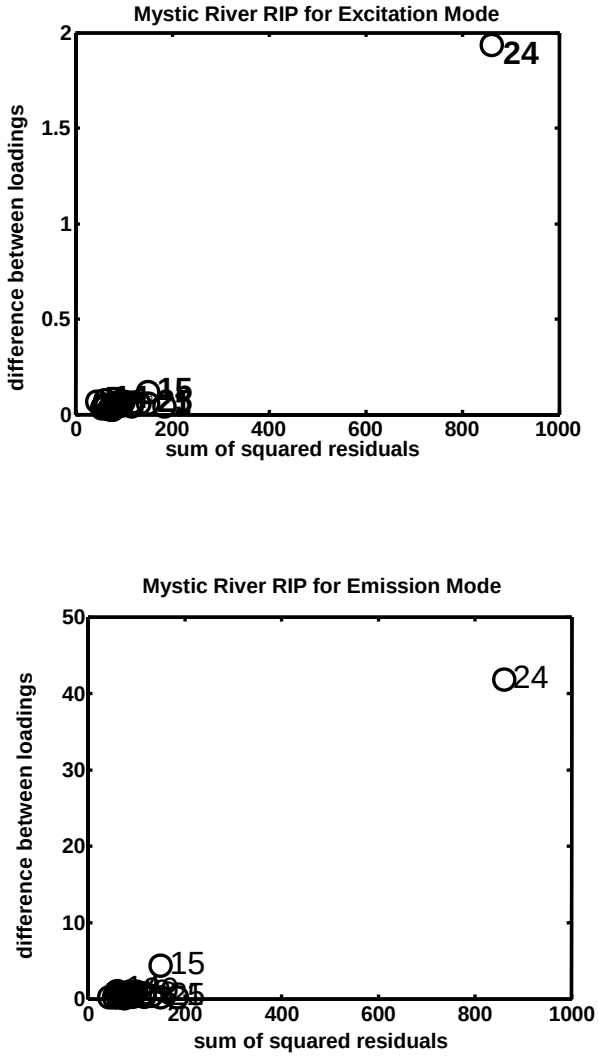

Figure S2 – Resample Influence Plots for Mystic River. Since sample 24 is a significant distance from the rest of the samples in both modes it can be considered as an outlier.

As can been seen above, sample 24 is the only sample that is not close to the identity line of the IMP. In addition, sample 24 is also far removed from the grouping of all other samples in the RIP. Therefore, sample 24 was removed from the data set. No outliers were found in the data from Mystic Lake or Chelsea River. One outlier was removed from the Mystic River and Boston Harbor sites. The outliers were from different dates.

Literature Cited

(1) Bro, R.; Kiers, H. A. L. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics* **2003**, *17* (5), 274-286.

(2) Riu, J.; Bro, R. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics Intell. Lab. Syst.* **2003**, *65* (1), 35-49.