Multi-way Analysis for Process Chemometrics with an Application to Water Quality Monitoring

NB Gallagher\textsuperscript{1}, BM Wise\textsuperscript{1}, D O'Sullivan\textsuperscript{1}, RT Roginski\textsuperscript{1}, AM Gilmore\textsuperscript{2}

\textsuperscript{1}Eigenvector Research, Inc. Manson, WA, USA.
\textsuperscript{2}Fluorescence Division, HORIBA Instruments Inc., Edison, NJ, USA.
Abstract

Excitation emission fluorescence is quickly becoming the method of choice for analysis of dissolved organic material in water streams because of compact instrumentation, ease of use, and it can take advantage of the unique properties of chemometrics methods such as parallel factor analysis (PARAFAC). In the example discussed here, the Horiba Aqualog Fluorometer and Datastream software interfaces with Eigenvector’s Solo software for PARAFAC analysis. A key aspect of the application is that Datastream implements a fully NIST traceable EEM while simultaneously making absorbance spectral measurements used to automatically correct the inner-filter effects thus providing a signal appropriate for PARAFAC analysis. Applications include monitoring drinking water treatment systems, wastewater streams, and environmental monitoring of water bodies and example analytes of interest include fulvic acid, humic acid and total organic carbon. This talk will contrast common chemometrics methods with PARAFAC and discuss the necessity for data pre-treatment methods. Finally, an example of a real-time application of Datastream interfaced with Solo for water quality monitoring will be demonstrated. The result is a simple to use application of advanced chemometrics models for process chemometrics.
Problem Statement

• Need for rapid and accurate quantification of dissolved organic matter (DOM) and total organic carbon (TOC) in drinking water

• Excitation-Emission Matrices (EEM)
  – Fluorescence signal useful for aromatic compounds

• Signal processing to
  – remove artifacts (e.g., Raleigh scatter) and
  – linearize the problem e.g., remove inner filter effects
  – provide data amenable to classical least squares (CLS) and parallel factor analysis (PARAFAC)
Automated Plant Monitoring

Measurement

The Aqualog®
Horiba patented technology
Dissolved Organic Matter

1. Fulvic Acids
   - Lower MW
   - Hydrophilic
   - Less Aromatic

2. Humic Acids
   - High MW
   - Hydrophobic
   - Highly Aromatic
   - Adsorbed by Coagulants

3. Amino Acids
   - Associated with Wastewater Effluent
   - Biopolymers

Tryptophan
Tyrosine
**Datastream™**

Interactive web interface to Solo_Predictor developed for Horiba Scientific: water quality monitoring.

Data from PARAFAC model and other sources.

Custom interface developed in partnership with MazamaScience.com. (Javascript, Angular and D3 visualization)

start demo
### Middlesex Water | Latest Sample

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Update</th>
<th>Raw</th>
<th>Settled</th>
<th>Finished</th>
<th>% Removal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqualog_TOC</td>
<td>2018-02-14</td>
<td>2.8, 2.8</td>
<td>2.8, 2.2</td>
<td>1.5, 1.5</td>
<td>47 %</td>
</tr>
<tr>
<td>A254</td>
<td>2018-02-14</td>
<td>0.078, 0.078</td>
<td>0.043, 0.033</td>
<td>0.020, 0.020</td>
<td>74 %</td>
</tr>
<tr>
<td>A680</td>
<td>2018-02-14</td>
<td>0.0020, 0.0021</td>
<td>0.0010, 0.00057</td>
<td>-0.000061, -0.00012</td>
<td>104 %</td>
</tr>
<tr>
<td>tot_fluor</td>
<td>2018-02-14</td>
<td>1.1e+4, 1.1e+4</td>
<td>4.0e+3, 4.0e+3</td>
<td>2.6e+3, 2.6e+3</td>
<td>76 %</td>
</tr>
<tr>
<td>humic_ind</td>
<td>2018-02-14</td>
<td>0.83, 0.82</td>
<td>0.80, 0.79</td>
<td>0.78, 0.78</td>
<td>5 %</td>
</tr>
<tr>
<td>fulvic</td>
<td>2018-02-14</td>
<td>0.015, 0.015</td>
<td>0.019, 0.019</td>
<td>0.020, 0.020</td>
<td>-37 %</td>
</tr>
<tr>
<td>humic</td>
<td>2018-02-14</td>
<td>0.013, 0.013</td>
<td>0.0093, 0.0092</td>
<td>0.0085, 0.0085</td>
<td>34 %</td>
</tr>
<tr>
<td>tryp</td>
<td>2018-02-14</td>
<td>0.0058, 0.0061</td>
<td>0.0067, 0.0070</td>
<td>0.0066, 0.0066</td>
<td>-11 %</td>
</tr>
<tr>
<td>h2f</td>
<td>2018-02-14</td>
<td>0.88, 0.87</td>
<td>0.48, 0.48</td>
<td>0.43, 0.42</td>
<td>52 %</td>
</tr>
<tr>
<td>h2fft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alkalinity</td>
<td>2018-02-11</td>
<td>46</td>
<td>40</td>
<td>46</td>
<td>0 %</td>
</tr>
<tr>
<td>Cl2</td>
<td>2018-02-11</td>
<td>4.3</td>
<td>1.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>THM</td>
<td>2018-02-11</td>
<td></td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOC</td>
<td>2018-02-11</td>
<td><strong>2.8</strong></td>
<td><strong>HI 1.1</strong></td>
<td><strong>LO 1.1</strong></td>
<td><strong>60 %</strong></td>
</tr>
<tr>
<td>UV254</td>
<td>2018-02-11</td>
<td>0.68</td>
<td>0.21</td>
<td>0.12</td>
<td>82 %</td>
</tr>
<tr>
<td>pH</td>
<td>2018-02-11</td>
<td>7.3</td>
<td>6.9</td>
<td>7.5</td>
<td>-2 %</td>
</tr>
<tr>
<td>Date</td>
<td>pH</td>
<td>Alkalinity (mg/l)</td>
<td>UV-254 (abs/5 cm)</td>
<td>TOC (mg/l)</td>
<td>SDSTHMFP (ug/l)</td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>-------------------</td>
<td>-------------------</td>
<td>------------</td>
<td>----------------</td>
</tr>
<tr>
<td>1/15/2018</td>
<td>7.12</td>
<td>44</td>
<td>0.821</td>
<td>3.42</td>
<td>222.74</td>
</tr>
<tr>
<td>1/17/2018</td>
<td>7.29</td>
<td>46</td>
<td>0.810</td>
<td>3.28</td>
<td></td>
</tr>
<tr>
<td>1/19/2018</td>
<td>7.33</td>
<td>50</td>
<td>0.747</td>
<td>3.11</td>
<td></td>
</tr>
<tr>
<td>1/21/2018</td>
<td>7.29</td>
<td>50</td>
<td>0.749</td>
<td>3.45</td>
<td></td>
</tr>
<tr>
<td>1/24/2018</td>
<td>7.19</td>
<td>38</td>
<td>1.365</td>
<td>6.76</td>
<td></td>
</tr>
<tr>
<td>1/26/2018</td>
<td>7.16</td>
<td>44</td>
<td>1.042</td>
<td>4.80</td>
<td></td>
</tr>
<tr>
<td>1/28/2018</td>
<td>7.28</td>
<td>56</td>
<td>0.765</td>
<td>3.60</td>
<td></td>
</tr>
<tr>
<td>1/30/2018</td>
<td>7.34</td>
<td>50</td>
<td>0.682</td>
<td>2.98</td>
<td></td>
</tr>
<tr>
<td>2/1/2018</td>
<td>7.40</td>
<td>58</td>
<td>0.626</td>
<td>2.76</td>
<td></td>
</tr>
<tr>
<td>2/3/2018</td>
<td>7.36</td>
<td>50</td>
<td>0.601</td>
<td>2.68</td>
<td></td>
</tr>
<tr>
<td>2/5/2018</td>
<td>7.26</td>
<td>44</td>
<td>0.895</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>2/7/2018</td>
<td>7.16</td>
<td>42</td>
<td>1.075</td>
<td>5.27</td>
<td></td>
</tr>
<tr>
<td>2/9/2018</td>
<td>7.27</td>
<td>38</td>
<td>0.966</td>
<td>4.39</td>
<td></td>
</tr>
<tr>
<td>2/11/2018</td>
<td>7.33</td>
<td>46</td>
<td>0.677</td>
<td>2.82</td>
<td></td>
</tr>
</tbody>
</table>
PARAFAC model estimates for Aqualog_TOC
Independent estimates of TOC
derived quantities

**Aqualog_TOC**

\[ a \times \frac{UV\ 254}{(Humic/Fulvic)} \]

\[
\begin{array}{cccccc}
a & b & c & d & e \\
0.0318 & & & & \\
\end{array}
\]

**Aqualog_A254**

\[ UV\ 254 \]

\[
\begin{array}{cccccc}
a & b & c & d & e \\
 & & & & \\
\end{array}
\]

**Aqualog_SUVA**

\[ 100 \times \frac{UV\ 254}{TOC} \]

\[
\begin{array}{cccccc}
a & b & c & d & e \\
 & & & & \\
\end{array}
\]

**Aqualog_THMFP**

\[ a \times TOC + b \times (Humic/Fulvic) + c \times Cl2 + d \times pH + e \times Alkalinity \]

\[
\begin{array}{cccccc}
a & b & c & d & e \\
26.51 & 22.48 & 0.782 & 11.70 & 0.883 \\
\end{array}
\]
Chemometrics

• Classical Least Squares (CLS) and Parallel Factor Analysis (PARAFAC)
  – model identification (parameterization) is different
    • CLS typically uses more parameters and is more flexible
    • PARAFAC more natural for EEMs, less prone to over-fitting, unique solution (true underlying factors)
  – model application (prediction) is similar
  – both can use non-negativity
  – both are linear wrt composition …
\[ \mathbf{X} = \mathbf{A}\mathbf{B}^T + \mathbf{E} \]

- **Excitation (nm)** • **Emission (nm)**
- **Row 1** • **Row 2**
- **Sample**
- **Matrix** \( I \times JK \)

**Classical Least Squares (CLS):**
Number of model parameters = \( RIJK \).

**Parallel Factor Analysis (PARAFAC):**
Number of model parameters = \( RI(J+K) \).

\[ \mathbf{X} \Rightarrow \mathbf{X} ; \quad \mathbf{X} = \mathbf{A}\left(\mathbf{B} \ast \mathbf{C}\right)^T + \mathbf{E} \]

- **Slab 1** • **Slab 2**
- **Sample**
- **Matrix** \( I \times J \times K \)

\( I = \) number of samples / objects,
\( J = \) number of excitation wavelengths,
\( K = \) number of emission wavelengths &
\( R = \) the number of factors / analytes.
\[ X = \sum_{r=1}^{R} a_r \otimes b_r + E \]

\[ X = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r + E \]
# Model Comparison

<table>
<thead>
<tr>
<th>Model Ambiguity</th>
<th>MCR/CLS</th>
<th>PARAFAC</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiplicative</td>
<td>yes</td>
<td>yes</td>
<td>Resolvable with constraints or standards</td>
</tr>
<tr>
<td>Permutation</td>
<td>yes</td>
<td>yes</td>
<td>trivial problem</td>
</tr>
<tr>
<td>Rotational</td>
<td>yes</td>
<td>no</td>
<td>data dependent: MCR requires selective data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PARAFAC requires tri-linearity</td>
</tr>
</tbody>
</table>

- MCR/CLS has more parameters, $RIJK$, and is therefore more flexible. Interpretation can be more difficult because excitation and emission modes are “convolved.”

- PARAFAC has fewer parameters, $RI(J+K)$, and is therefore less prone to overfit. Interpretation can be easier because the excitation and emission modes are modeled “independently.” Uniqueness during model identification
Rayleigh Scatter
Inner Filter Effects are attenuation of excitation and/or emission signal due to dissolved analytes. The ability to make this correction is enabled by simultaneous measurement of an absorbance spectrum.

If 50% of each then

$$\int_{\lambda_{ex}, \lambda_{em}}^{IFE} = 10^{0.5(A(\lambda_{ex})+A(\lambda_{em}))}$$

$A(\lambda_{ex})$ is the primary absorbance

$A(\lambda_{em})$ is the secondary absorbance
Accounting for inner filter effects results in a signal linear with concentration.
Rapidly Assess DOM in Water Streams

C1: Fulvic acids
Lo MW and Aromaticity

C2: Humic acids
Hi MW and Aromaticity

C. c3
Protein like
Conclusions

- Aqualog provides fast EEMs used for online process analysis
- Fluorescence signal useful for aromatic compounds in water streams
  - humic acids, fulvic acids, proteins (e.g., tryptophan), DOM
- CLS and PARAFAC models were used for quantification
  - CLS has more parameters: flexible but more prone to overfit
  - PARAFAC has fewer parameters: more natural for EEMs
- Signal processing is necessary to enable the models
  - Raleigh Scatter and Inner Filter Effects
- Additional applications
  - organic membrane fouling compounds
  - ID & quantify algae: bluegreens, diatoms and chlorophytes,
  - polycyclic aromatic hydrocarbons
  - wine quality (phenolics): sulfites, spoilage