EIGENVECTOR RESEARCH INCORPORATED
Research, Training, and Software

196 Hyacinth Road
Manson, WA 98831
www.Eigenvector.com

PARAFAC for Analysis of Fluorescence EEM Data

Rasmus Bro and Neal B. Gallagher

Key words: Variable weighting, pattern recognition, generalized least squares

Introduction: While principal component analysis is useful for analysis of two-way data, parallel factor analysis (PARAFAC) is the method of choice for three-way data.[1] For example, two-way data might consist of N measurements (e.g., spectral absorbance at different wavelengths) on M objects (or samples) with the data collected into a matrix $\mathbf{X}_{M\times N}$. In contrast, EEM data consists of measurements of fluorescence at P excitation wavelengths and N emission wavelengths for each of M samples with the data collected into a three-way "data cube" or "box" $\underline{\mathbf{X}}_{M\times N\times P}$. Although PARAFAC can be extended to general multi-way data, the focus here is three-way EEM data.

One way to write the PARAFAC model is

$$\underline{\mathbf{X}}_p = \mathbf{A}\mathbf{D}_p\mathbf{B}^T + \underline{\mathbf{E}}_p \quad p = 1,...,P \qquad (1)$$

where $\underline{\mathbf{X}}_p$ is the emission spectrum for the p^{th} excitation. For a model with K factors, $\mathbf{A}$ is M by K and corresponds to loadings in the sample mode. This is what we normally call the score matrix. The matrix $\mathbf{B}$ is N by K and holds the loadings in the emission mode, $\mathbf{D}_p$ is a K by K diagonal matrix. It's elements are the p^{th} row of the P by K loading matrix $\mathbf{C}$ which contains the excitation loadings. Hence, at any given excitation wavelength p, the model of the measured emission spectra, $\underline{\mathbf{X}}_p$, is given by the same scores, $\mathbf{A}$, and the same emission loadings, $\mathbf{B}$, only each component is weighted by specific excitation loadings as defined in the diagonal matrix $\mathbf{D}_p$. The matrix $\underline{\mathbf{E}}_p$ is the model residuals for the p^{th} sample.

The real 'trick' in PARAFAC is that, unlike PCA, PARAFAC is unique. Hence, if the data follows a PARAFAC model, then the emission loadings are not just abstract orthogonal emission profiles as would be the case in PCA. Instead, the loadings are actually estimates of the real emission spectra of the real fluorophores.

Experimental: A Perkin-Elmer LS50 B fluorescence spectrometer was used to measure fluorescence landscapes using excitation wavelengths between 200-350 nm with 5 nm intervals. The emission wavelength range was 200-750 nm. Excitation and emission monochromator slit widths were set to 5 nm, respectively. Scan speed was 1500 nm/min. Measurements were made on mixtures of four flourophores (hydroquinone, tryptophan, phenylalanine and dopa) at known concentrations (ppm M) giving a data set that was 27 samples by 121 emission wavelengths by 24 excitation wavelengths [The example uses the `dorrit` data set.[2]]

Results and Discussion:

Figure 1 shows an EEM landscape for a single sample corresponding to DOPA at 55 ppm. Equation 1 represents a tri-linear model, and ideally all sources contributing to the measured signal would follow this model. The two major peaks in Figure 1 likely have a response that is fairly close to tri-linear however, the Rayleigh scattering indicated by the arrow (corresponding to the line where excitation equals emission) does not adhere to the tri-linear model.

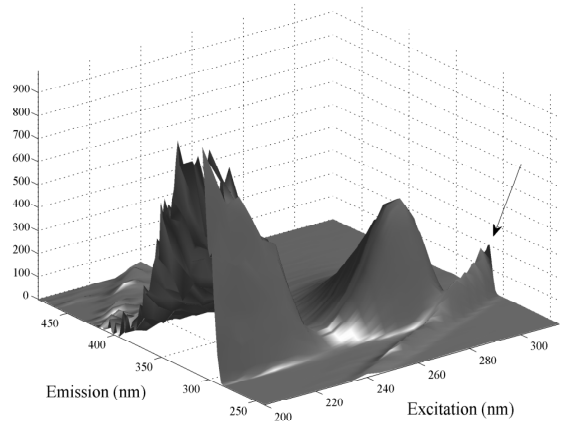

Figure 1: EEM of pure DOPA at 55 ppm. The arrow indicates a ridge associated with Rayleigh scattering.

Additionally, the top of the large peak is saturated and will not follow the PARAFAC model. As a result, these regions are treated as 'missing data' by replacing

196 Hyacinth Road
Manson, WA 98831
www.Eigenvector.com

their entries with NaN (Not-a-Number). Regions where emission is greater than excitation, and strictly not included in the Rayleigh scatter, are set to zero as expected from the physics of the EEM measurement. This is done using the FLUCUT function:

```
Xnew = flucut(X,20,[20 20],NaN,NaN,0,0);
```

In this example, entries below emission = excitation - 20 nm were set to zero, while a band of 20 nm above and below emission = excitation were set to NaN (missing). The result for DOPA at 55 ppm is shown in Figure 2. Note that the secondary Rayleigh scattering was not accounted for in this example.

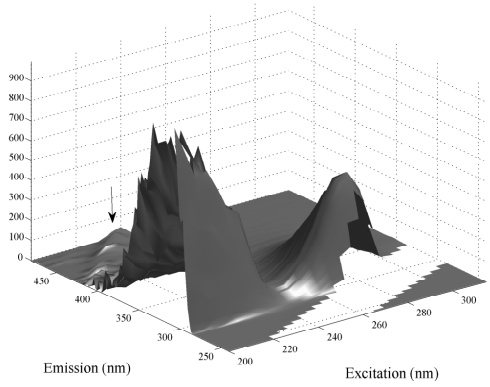

Figure 2: EEM of pure DOPA at 55 ppm with Rayleigh scattering set to missing.

Figure 3 shows that the contributions in mode 1 for factor 2 (column 2 of **A**) correspond to the known DOPA concentration. The excitation (**C**) and emission (**B**) profiles are shown for the four factors in Figures 4 and 5 respectively.

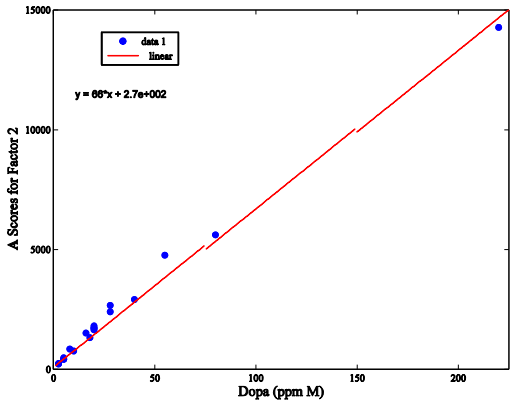

Figure 3: Scores on factor 2 versus DOPA concentration.

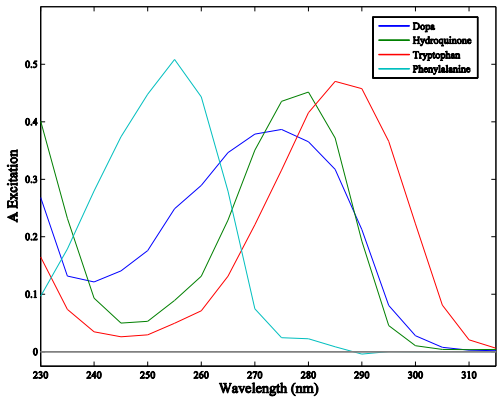

Figure 4: Excitation profiles (loadings A).

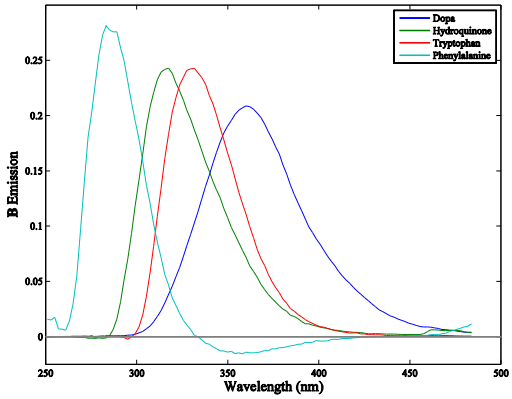

Figure 5: Emission profiles (loadings B).

Conclusions: The tri-linear PARAFAC model can be easily applied to EEM data to provide interpretable results. However, Rayleigh scattering and signal saturation must be accounted for to avoid artifacts in the analysis.

References:

[1] Smilde, A., Bro, R., Geladi, P., "Multi–way Analysis with Applications in the Chemical Sciences," John Wiley & Sons, New York, NY (2004).

[2] J. Riu, R. Bro "Jack-knife estimation of standard errors and outlier detection in PARAFAC models," *Chemometr. Intell. Lab.* 2003; **65**(1): 35–49.