Neural Networks vs. Whittaker Smoothing: Advanced Techniques for Scattering Signal Removal in 3D Fluorescence spectra

Aleksandr S. Zakuskin
Department of Chemistry
Lomonosov Moscow State University
Moscow, Russia
ale-zakuskin@laser.chem.msu.ru

Ivan N. Krylov
Department of Chemistry
Lomonosov Moscow State University
Moscow, Russia
ikrylov@laser.chem.msu.ru

Timur A. Labutin
Department of Chemistry
Lomonosov Moscow State University
Moscow, Russia
timurla@laser.chem.msu.ru

Abstract

Fluorescence excitation emission matrices (EEMs) have a trilinear structure, aligning perfectly with the tensor rank decomposition, PARAFAC. Consequently, PARAFAC has become essential for extracting information from freshwater EEMs, pinpointing individual fluorophore groups, and tracking their behaviour across diverse environment. However, EEMs of seawater, with typically low organic matter, are often dominated by Rayleigh and Raman scattering, which deviates from the trilinear model. Traditional one-dimensional interpolation to eliminate these interferences varies in outcome based on its matrix application direction and struggles with noisy data. Our proposed techniques, employing Whittaker smoothing and CNN, effectively eliminate scattering signals, even in noise-rich scenarios. Notably, CNN adeptly preserves the overall EEM shape across various sizes and dimensions, establishing itself as an optimal choice for interpolating scattering zones in EEMs of organic matter-deficient freshwaters.

1 Introduction

The origins, transformations, and conservation mechanisms of dissolved organic matter (DOM) remain pivotal questions within marine and aquatic sciences. The systematic study of these properties augments our understanding of environmental dynamics and the intricate interplay between carbon and nitrogen cycles in watersheds [1]. Due to the highly intricate composition of DOM in natural waters and the overlapping wide-band spectra, fluorescence signals cannot be attributed to individual compound. Hence, researchers typically focus on groups of DOM fluorophores – sets of compounds associated with distinct fluorescence bands, which in turn relate to the origin and transformation of organic matter in waters [2]. The fluorescence excitation-emission matrices (EEMs), which capture fluorescence signals across a multitude of excitation/emission wavelength pairs for each sample, offer profound insights into DOM. The trilinear structure of EEM data aligns seamlessly with the tensor rank decomposition known as PARAFAC [3]. PARAFAC has risen as a pivotal tool for identifying individual components across varied environmental backdrops [4]. This synergistic combination of EEM and PARAFAC is used for a broad range of freshwater studies [5], and extending to evaluations of water treatment efficacy and quality control [6, 7].

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

Certain sections of EEMs can pose challenges in PARAFAC modelling because they do not adhere to a trilinear structure. The two primary sources of interference are Rayleigh and Raman scattering [8]. Unfortunately, signals from both Rayleigh and Raman scattering are often prevalent in EEMs of seawater, which typically have a low DOM content. Various techniques have been tested to eliminate these inferences, including down-weighting of the scatter region (MILES), specific scatter modelling, subtraction of a standard, application of constraints during decomposition, insertion of missing values, or setting zeros outside the data region [3]. Of these, interpolation generally delivers the highest efficiency [9]. The commonly used one-dimensional interpolation yields different results depending on whether it is conducted row-wise or column-wise on the matrix. Moreover, in instances where the noise level is high, which is often the case in natural waters, there is no guarantee that the fluorescence signal neighbouring the scattering region will be strictly monotonic [10]. Thus, we have focused our study on the implementation of multidimensional strategies to effectively purge interference from scattering prior to PARAFAC modelling. We introduce two techniques, namely, two-dimensional Whittaker smoothing [11, 12] and a purpose-built CNN model to rectify the signal in areas affected by scattering.

2 Datasets

Synthetic Dataset 1 based on OpenFluor We generated two synthetic datasets using known "ground truth" fluorophore spectra and their concentrations to compare the performance of the studied methods. The first dataset was created for the selection, construction, and training of the optimal neural network. It was derived from the PARAFAC decompositions of real spectra gathered in the OpenFluor database [13]. From 230 sets of fluorophores with randomly generated concentration values, we obtained a total of 100,000 spectra: 80,000 for the training set, 10,000 for the validation set, and 10,000 for the test set.

The scattering bands in question are: 1st order Rayleigh scattering band ($\lambda_{\mathrm{em}} = \lambda_{\mathrm{ex}}$), 1st order Raman scattering band ($\lambda_{\mathrm{em}} = (1/\lambda_{\mathrm{ex}} - \Delta_{\mathrm{Raman}})^{-1}$), 2nd order Rayleigh scattering band ($\lambda_{\mathrm{em}} = 2\lambda_{\mathrm{ex}}$), and 2nd order Raman scattering band ($\lambda_{\mathrm{em}} = 2(1/\lambda_{\mathrm{ex}} - \Delta_{\mathrm{Raman}})^{-1}$).

Synthetic Dataset 2 Set of spectra was also obtained for which it was possible to perform PARAFAC. The loadings for the simulated datasets were constructed from Gaussian peak functions centred across the typical emission and excitation range of DOM. The dataset consists of 64 samples. Scores were calculated using an orthogonal Latin hypercube design. The scattering signal was added in the same manner as for the first dataset. We investigated the sensitivity of the proposed scatter handling methods to noise, adding varying levels of specially introduced noise to the spectra, with ratios ranging from 10^{-3} to $5 \cdot 10^{-1}$.

"Fluordata" dataset An experimental fluorescence dataset generated and measured by Åsmund Rinnan and Jordi Riu has been chosen to demonstrate the differences between interpolation methods [14] (class 4 contains 5 fluorophores). We have added the similar scattering band and vary the noise level.

3 Methods

Whittaker smoother Whittaker smoothing [15] requires the surface to be sampled on a grid. Given a vector of values $\mathbf{z}$ measured at uniform intervals, Whittaker smoother $\hat{\mathbf{z}}$ is computed by minimising $|\mathbf{z} - \hat{\mathbf{z}}|^2 + \lambda |\mathbf{D}\hat{\mathbf{z}}|^2$ where the matrix $\mathbf{D}$ is constructed in such a way that multiplying it by a vector results in a vector of differences of a given order between the points of the vector. The penalty weight λ strikes a balance between smoothness and fidelity. We unfold the excitation-emission matrix $\mathbf{F}$ ($F_{i,j} = F(\lambda_i^{\mathrm{em}}, \lambda_j^{\mathrm{ex}})$) into a vector $\mathbf{z}$ ($z_{i+j \cdot N} = F_{i,j}$ for an N-row matrix) and then construct the appropriate two-dimensional penalty using Kronecker products: $\mathbf{D} = \begin{pmatrix} \mathbf{D}_{\mathrm{x}} \otimes \mathbf{I} \\ \mathbf{I} \otimes \mathbf{D}_{\mathrm{y}} \end{pmatrix}$. For non-uniform grids, we solve the Vandermonde system, finding the coefficients required to estimate the kth order derivative from each $k+1$ successive points of the grid [16]. We developed a sequential algorithm to interpolate surfaces sampled on arbitrary grids, adhering to smoothness penalties of variable orders, utilising Whittaker smoothing generalised for two-dimensional spectral data.

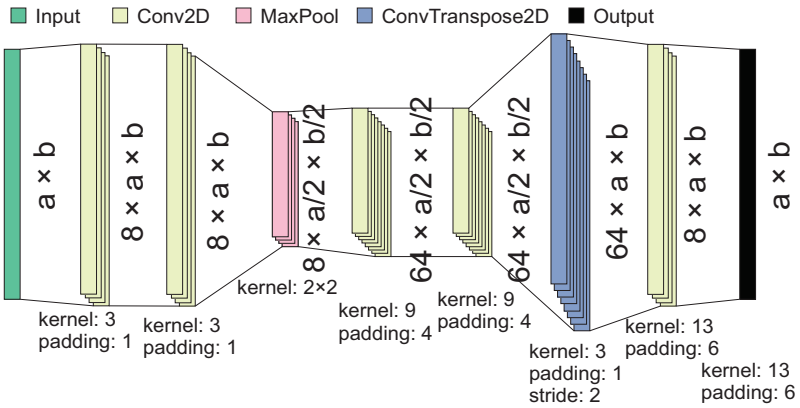


Figure 1: Schematic representation of the final CNN architecture

In this work, difference orders of 1 and 2 were combined in different proportions to choose the best shape of the estimated fluorescence signal. Since theoretical selection of penalty weight is not possible, we have performed their optimisation on the synthetic dataset, minimising the root-mean-squared error of signal reconstruction.

CNN description The resulting CNN model (Figure 1) consists of three blocks of 2 convolutional layers each with `MaxPool` layer between the first two blocks and `ConvTranspose` layer between 2nd and 3rd blocks. Each layer was followed by `ReLU` activation. Adam optimiser and `ReduceLROnPlateau` scheduler were used. Training was done by minibatches of 16 EEMs at a time. Since dimensions of EEMs are different, they were padded by zeros on each side to the maximum dimension in a given minibatch, then padded again to make both dimensions even as required by pooling layer in the architecture. Therefore, the model does not require fixed dimensionality of input data for either training or inference.

The interpolation methods were compared based on the two performance metrics. For synthetic datasets, where "ground truth" EEMs are known by construction, the resulting EEMs were compared to the original ones, using the root mean squared error (RMSE) of reconstruction of the ground truth fluorescence signal as the metric. The second synthetic dataset as well as the "fluordata" dataset were decomposed using non-negative PARAFAC, and the estimated loadings were compared to the ground truth values. The performance metric employed is the Tucker's congruence coefficient (cosine similarity, TCC) between the "ground truth" excitation and emission spectra and the excitation and emission loadings estimated by PARAFAC after EEMs correction. The PARAFAC decomposition was performed in the `albatross` R package.

4 Results

We began our consideration with the Synthetic dataset 1 (SD1). It would be straightforward to evaluate the performance of our technique in terms of RMSE for the entire spectrum. Above all, it's worth noting the tremendous advantage of the CNN in terms of performance, 0.024 seconds per spectrum, compared to 1.464 seconds when using Whittaker smoother. We identified a statistically significant (via *t*-test) reduction in RMSE for CNN (Figure 2, left). Meanwhile, it's essential to highlight the distinct features of using both approaches. At low resolution (Figure 2, top row), the CNN effectively removes the scattering signal, which consists of a single pixel, but also significantly 'blurs' the fluorescence signal to adjacent pixels. In contrast, Whittaker smoother works wonderfully in such cases, as interpolation effectively occurs within a single pixel's boundaries. In the case of high resolution (Figure 2, bottom row), where scattering intersects with a large fluorescence signal area, smoothing doesn't fully interpolate the signal and beyond this area, which apparently relates to the influence of neighbouring areas in two-dimensional interpolation. With CNN, the blurring of the signal within a small number of pixels doesn't have a significant impact. Based on this, we suggested to processes the entire image as a whole (CNN) and also to replace part of the signal within the scattering signal's presence (CNN2).

Since the main task was the improvement of PARAFAC decomposition, we have compared the obtained PARAFAC components for Synthetic Dataset 2 (SD2) after scattering signal removal. TCC between ground truth loadings and matching PARAFAC-estimated loadings (averaged over all PARAFAC components) again demonstrates the ultimate performance of both techniques with slight advantage of Whittaker (Figure 3b). But it also should be noted that the CNN keeps the advantage of Whittaker methods to recover the true excitation and emission spectra of fluorophores in presence of noise. In general, the RMSE is low across all noise level, while RMSE grew in two times for the highest noise level in case of Whittaker. Nevertheless, both approaches again clearly demonstrate the ability of the techniques to recover the shape of the fluorescence signal itself. This issue might be caused by the use of data with an extremely asymmetric excitation and emission scale (35×351). This dataset was used for the convenience of comparison with previously obtained data for other technique for scattering handling.

Lastly, when examining the real data from the Åsmund dataset (as shown in the Figure 3a), we find that the neural network takes the lead when its results applied solely to the scattering signal's range. Nonetheless, using the entire EEM matrix as an option also yields satisfactory results, even at a noise

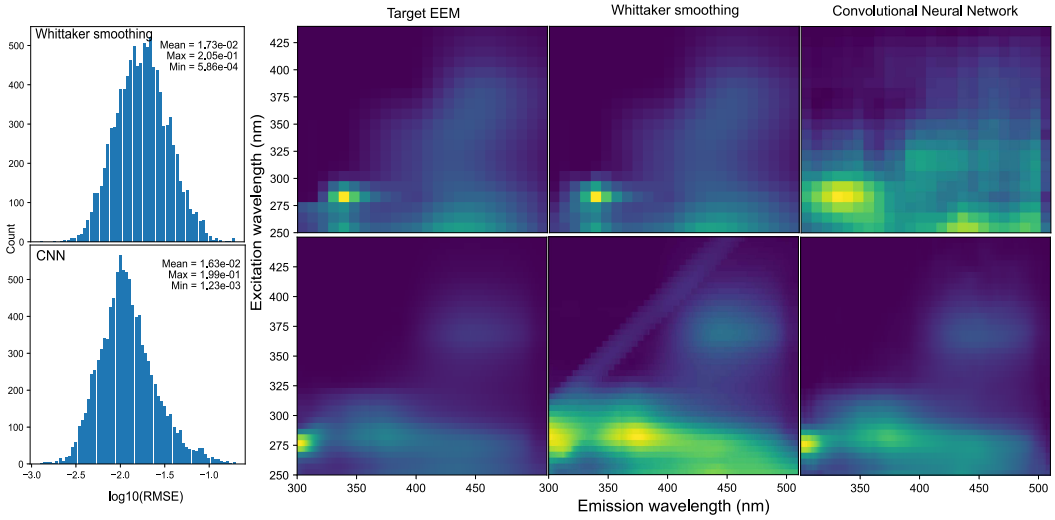
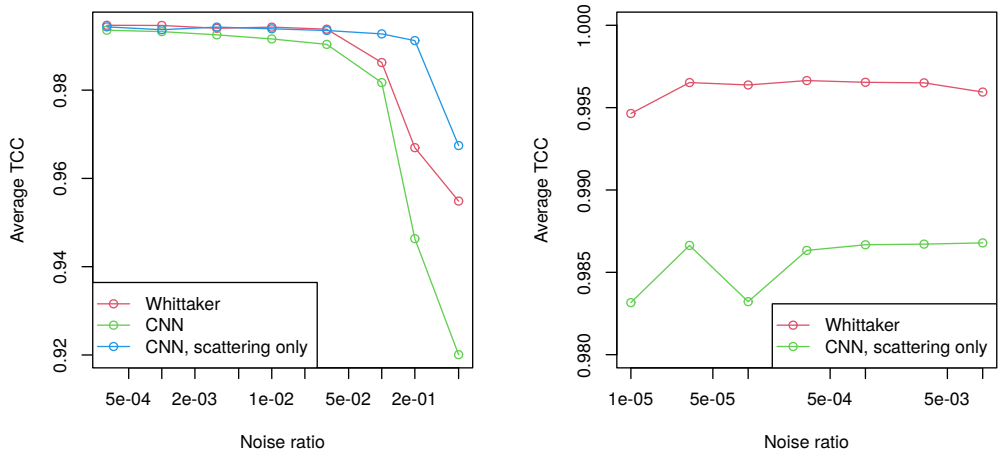


Figure 2: The best and the worst results obtained by both methods for samples from SD1, and the distribution of their recovery error.



(a) The "fluordata" dataset.

(b) The synthetic dataset 2.

Figure 3: Average Tucker's congruence coefficient between the ground truth loadings and the estimated loadings for the two datasets for which PARAFAC was feasible.

level of 0.2. If we consider the components resulting from decomposition, it's worth noting the high resilience of PARAFAC to noise in the absence of interference from scattering signals. The most challenging component to reconstruct – the fluorescence of tyrosine, which considerably overlaps with the scattering signals – is reliably predicted at high noise levels with the use of CNN2 only.

5 Conclusions

Thus, we can conclude that both Whittaker smoothing and CNN methods offer high performance for scattering signal removal across various signals, even with high levels of noise. However, the most optimal approach involves replacing the scattering area with a signal processed using CNN, enabling signal processing up to a 0.5 noise-to-signal ratio. Furthermore, CNN preserves the general shape of the EEM, regardless of its dimensionality and size, in the most accurate manner. This positions CNN as an ideal candidate for interpolating scattering areas in EEMs of fresh waters with low levels of DOM, and consequently, a low signal-to-noise ratio. These observations were exemplified using the real "fluordata" dataset, which was obtained from a mixture of prevalent natural fluorophores.

Acknowledgments and Disclosure of Funding

The work was supported by the Fellowship from Non-commercial Foundation for the Advancement of Science and Education INTELLECT.

References

[1] R. Jaffé, D. McKnight, N. Maie, R. Cory, W. H. McDowell, and J. L. Campbell. Spatial and temporal variations in DOM composition in ecosystems: The importance of long-term monitoring of optical properties. *Journal of Geophysical Research: Biogeosciences*, 113(G4), December 2008. ISSN 01480227. doi: 10.1029/2008JG000683. URL `http://doi.wiley.com/10.1029/2008JG000683`.

[2] Rose M. Cory and Diane M. McKnight. Fluorescence spectroscopy reveals ubiquitous presence of oxidized and reduced quinones in dissolved organic matter. *Environmental science & technology*, 39(21):8142–8149, November 2005. ISSN 0013-936X. doi: 10.1021/es0506962. URL `https://doi.org/10.1021/es0506962`.

[3] Saioa Elcoroaristizabal, Rasmus Bro, Jose Antonio García, and Lucio Alonso. PARAFAC models of fluorescence data with scattering: A comparative study. *Chemometrics and Intelligent Laboratory Systems*, 142:124–130, March 2015. ISSN 01697439. doi: 10.1016/j.chemolab.2015.01.017. URL `http://linkinghub.elsevier.com/retrieve/pii/S0169743915000283`.

[4] Liyang Yang, Jin Hur, and Wane Zhuang. Occurrence and behaviors of fluorescence EEM-PARAFAC components in drinking water and wastewater treatment systems and their applications: a review. *Environmental Science and Pollution Research*, 22(9):6500–6510, May 2015. ISSN 0944-1344, 1614-7499. doi: 10.1007/s11356-015-4214-3. URL `https://link.springer.com/article/10.1007/s11356-015-4214-3`.

[5] Rudolf Jaffé, Kaelin M. Cawley, and Youhei Yamashita. Applications of Excitation Emission Matrix Fluorescence with Parallel Factor Analysis (EEM-PARAFAC) in Assessing Environmental Dynamics of Natural Dissolved Organic Matter (DOM) in Aquatic Environments: A Review. In Fernando Rosario-Ortiz, editor, *ACS Symposium Series*, volume 1160, pages 27–73. American Chemical Society, Washington, DC, January 2014. ISBN 978-0-8412-2951-8 978-0-8412-2949-5. doi: 10.1021/bk-2014-1160.ch003. URL `https://pubs.acs.org/doi/abs/10.1021/bk-2014-1160.ch003`.

[6] Iván Sciscenko, Antonio Arques, Pau Micó, Margarita Mora, and Sara García-Ballesteros. Emerging applications of EEM-PARAFAC for water treatment: a concise review. *Chemical Engineering Journal Advances*, 10:100286, May 2022. ISSN 2666-8211. doi: 10.1016/j.ceja.2022.100286. URL `https://www.sciencedirect.com/science/article/pii/S2666821122000473`.

[7] Stephanie K. L. Ishii and Treavor H. Boyer. Behavior of Reoccurring PARAFAC Components in Fluorescent Dissolved Organic Matter in Natural and Engineered Systems: A Critical Review. *Environmental Science & Technology*, 46(4):2006–2017, February 2012. ISSN 0013-936X. doi: 10.1021/es2043504. URL https://doi.org/10.1021/es2043504. Publisher: American Chemical Society.

[8] Åsmund Rinnan and Charlotte M. Andersen. Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation–emission data. *Chemometrics and Intelligent Laboratory Systems*, 76(1):91–99, March 2005. ISSN 0169-7439. doi: 10.1016/j.chemolab.2004.09.009. URL http://www.sciencedirect.com/science/article/pii/S0169743904002242.

[9] Kathleen R. Murphy, Colin A. Stedmon, Daniel Graeber, and Rasmus Bro. Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Analytical Methods*, 5(23):6557, 2013. ISSN 1759-9660, 1759-9679. doi: 10.1039/c3ay41160e. URL http://xlink.rsc.org/?DOI=c3ay41160e.

[10] Anastasia N. Drozdova, Ivan N. Krylov, Andrey A. Nedospasov, Elena G. Arashkevich, and Timur A. Labutin. Fluorescent signatures of autochthonous dissolved organic matter production in Siberian shelf seas. *Frontiers in Marine Science*, 9:872557, 2022. ISSN 2296-7745. URL https://www.frontiersin.org/articles/10.3389/fmars.2022.872557.

[11] Clement Atzberger and Paul H. C. Eilers. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. *International Journal of Remote Sensing*, 32(13):3689–3709, July 2011. ISSN 0143-1161. doi: 10.1080/01431161003762405. URL https://doi.org/10.1080/01431161003762405. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431161003762405.

[12] Huazhou Chen, Wu Ai, Quanxi Feng, Zhen Jia, and Qiqing Song. FT-NIR spectroscopy and Whittaker smoother applied to joint analysis of duel-components for corn. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 118:752–759, January 2014. ISSN 1386-1425. doi: 10.1016/j.saa.2013.09.065. URL https://www.sciencedirect.com/science/article/pii/S138614251301072X.

[13] Kathleen R. Murphy, Colin A. Stedmon, Philip Wenig, and Rasmus Bro. OpenFluor– an online spectral library of auto-fluorescence by organic compounds in the environment. *Analytical Methods*, 6(3):658–661, 2014. ISSN 1759-9660, 1759-9679. doi: 10.1039/C3AY41935E. URL https://pubs.rsc.org/en/content/articlelanding/2014/ay/c3ay41935e.

[14] Rasmus Bro, Åsmund Rinnan, and Nicolaas (Klaas) M. Faber. Standard error of prediction for multilinear PLS: 2. Practical implementation in fluorescence spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 75(1):69–76, January 2005. ISSN 0169-7439. doi: 10.1016/j.chemolab.2004.04.014. URL https://www.sciencedirect.com/science/article/pii/S0169743904001236.

[15] Paul H. C. Eilers. A Perfect Smoother. *Analytical Chemistry*, 75(14):3631–3636, July 2003. ISSN 0003-2700. doi: 10.1021/ac034173t. URL https://doi.org/10.1021/ac034173t. Publisher: American Chemical Society.

[16] Bengt Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of Computation*, 51(184):699–706, 1988. ISSN 0025-5718, 1088-6842. doi: 10.1090/S0025-5718-1988-0935077-0. URL https://www.ams.org/mcom/1988-51-184/S0025-5718-1988-0935077-0/.