Ahmad Mohamad (Orcid ID: 0000-0001-5127-5707)
Vitale Raffaele (Orcid ID: 0000-0002-7497-1673)
Cocchi Marina (Orcid ID: 0000-0001-8764-4981)
Ruckebusch Cyril (Orcid ID: 0000-0001-8120-4133)

Weighted multivariate curve resolution – alternating least squares based on sample relevance

M. Ahmad[1,2*], R. Vitale[1], M. Cocchi[2], C. Ruckebusch[1]

[1] Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France.
[2] Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Modena, Italy.

Keywords: Weighted least squares, Multivariate curve resolution – alternating least squares (MCR-ALS), Essential information, Spectral pixels, Sample selection
Corresponding author: m.ahmad@live.nl

Abstract

Alternating least squares, within the multivariate curve resolution framework has seen a lot of practical applications and shows its distinction with its relatively simple and flexible implementation. However, the limitations of least squares should be considered carefully when deviating from the standard assumed data structure. Within this work we highlight the effects of noise in the presence of minor components, and we propose a novel weighting scheme within the weighted multivariate curve-resolution-alternating least squares framework, to resolve it. Two simulated and one Raman imaging case is investigated, by comparing the novel methodology against standard multivariate curve resolution-alternating least squares and essential spectral pixel selection. A trade-off is observed between current methods, while the novel weighting scheme demonstrates a balance, where the benefits of the previous two methods are retained.

1. Introduction

Multivariate curve resolution (MCR) is a methodology with its fingers in many fields [1]. Its ability to resolve unknown mixtures, combined with simple to understand algorithms and interpretable results, makes it a highly performant method. MCR is aimed at resolving the most linearly dissimilar sources of variances (which are assumed to be the purest) underlying bilinear data. One of the most utilized MCR algorithms is multivariate curve resolution –

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/cem.3478

alternating least squares (MCR-ALS). MCR-ALS has been proven to be effective in many practical scenarios however, within the ALS procedure, some of the well-known limitations of least squares approaches must be considered. One of these limitations in particular regards to the presence of non-independent and -identically distributed (non-iid) noise. To cope with this, a weighted MCR-ALS algorithm has been developed by Wentzell et al. based on maximum likelihood projections and applied to different types of data [2, 3]. Another limitation relates to the so-called "black-hole" effect as pointed out recently by Vitale et al. [4]. This issue is connected to the leverage that some data points may have in the non-negative least squares (NNLS) calculation. In MCR, single data points that are very far from the data cloud are potentially the purest ones. However, when utilising MCR-ALS, their leverage might become too low for guaranteeing the correct solution when the number of data points for other components is very large. Even starting from the most favourable initial estimates (i.e., the true pure profiles), the solution will in such a situation iteratively move closer to the centre of the data cloud. A solution to overcome this black-hole effect and improve the accuracy of the MCR-ALS output is sample selection, and an efficient way to do so is by selecting essential samples based on a convex hull criterium [5, 6, 7]. Examples of applications to hyperspectral imaging data showed that it is possible to recover similar or sometimes better solutions, with respect to standard MCR-ALS [8, 9, 10]. However, when noise comes into play, selecting too few samples can at some point decrease the stability of the model, as the number of points to properly estimate its parameters is greatly reduced [11]. In practical situations, there is a trade-off to be found, as reducing the data down to its most essential information will increase the variance of the estimated parameters while utilising the entire data set might reduce the accuracy of said parameters, as observed in the aforementioned black-hole effect.

In this short communication, we propose a weighted MCR-ALS methodology to balance this trade-off. To put it in perspective, the selection of samples based on essential information can be seen as the most extreme form of weighted analysis i.e., weighting one, essential samples, and zero, others. Here we propose a weighting scheme where sample weights are determined based on their relevance towards the MCR solution. To this aim, convex peeling [12] of the data set is performed i.e., repeated convex hull calculations, pruning the data, layer by layer until no points remain. For each layer, samples receive the same weight, with weights decreasing for the consecutive layers. A comparison is made between standard MCR-ALS, weighted MCR-ALS with weights encoding ESP selection and weighted MCR-ALS with weights encoding convex peeling, applied on three different data sets, two simulated and one real.

2. Methods

2.1 Weighted MCR-ALS

For the sake of brevity, MCR-ALS will not be detailed, we refer to de Juan et al. [1]. In this work, a modified ALS framework is formulated where, instead of applying a standard non-negative least squares approach to estimate the concentration and spectral profiles, a weighted version of the fast NNLS algorithm, developed by Bro et al. [13] is applied. The main differences with respect to standard MCR-ALS are presented in Eq. 1-5:

(Eq. 1) $$\boldsymbol{D} = \widehat{\boldsymbol{C}}\widehat{\boldsymbol{S}}^T + \boldsymbol{E}$$

Weighted MCR-ALS

(Eq. 2) $$\widehat{\boldsymbol{C}} = \boldsymbol{D}\boldsymbol{S}_{ini}(\boldsymbol{S}_{ini}^{T}\boldsymbol{S}_{ini})^{-1}$$

↓

(Eq. 3) $$\widehat{\boldsymbol{S}}^{T} = \left(\widehat{\boldsymbol{C}}^{T}\boldsymbol{W}_{c}\widehat{\boldsymbol{C}}\right)^{-1}\widehat{\boldsymbol{C}}^{T}\boldsymbol{W}_{c}\boldsymbol{D}$$

↓

(Eq. 4) $$\widehat{\boldsymbol{C}} = \boldsymbol{D}\widehat{\boldsymbol{S}}\left(\widehat{\boldsymbol{S}}^{T}\widehat{\boldsymbol{S}}\right)^{-1}$$

↓↑ (alternating)

(Eq. 5) $$\widehat{\boldsymbol{S}}^{T} = \left(\widehat{\boldsymbol{C}}^{T}\boldsymbol{W}_{c}\widehat{\boldsymbol{C}}\right)^{-1}\widehat{\boldsymbol{C}}^{T}\boldsymbol{W}_{c}\boldsymbol{D}$$

Where $\boldsymbol{D}$ $(I \times J)$ is the data matrix with I samples and J spectral channels, $\widehat{\boldsymbol{C}}$ $(I \times m)$ contains the estimated concentration profiles for m components, $\widehat{\boldsymbol{S}}$ $(J \times m)$ carries the estimated spectral profiles, $\boldsymbol{E}$ $(I \times J)$ is the model error matrix, $\boldsymbol{W}_c(I \times I)$ the weighting matrix for $\widehat{\boldsymbol{C}}$, and the spectra used as initial estimates are denoted as $\boldsymbol{S}_{ini}^{T}$. $\boldsymbol{W}_c$ is a square matrix with its diagonal containing the weights of all samples, and zeros elsewhere.

2.2 Weighting scheme

The samples are weighted according to their relevance to the MCR solution. In this work, we first used essential spectral points (ESPs) as the basis for the weighting scheme. ESPs can be found by taking the points along the convex hull of the normalized scores of the data set within its principal component subspace (PC-space) [5, 6, 7]. ESPs carry all the spectral information required for an accurate MCR resolution, and their selection reduces the data set considerably, without losing any essential information [5, 6, 7]. ESP selection can be encoded in $\boldsymbol{W}_c$, with weights equal to one, for ESPs, and weights equal to zero, for non-ESPs. However, this is the most extreme form of weighting.

We extended this approach by applying convex peeling [12], where each peel, l, is considered a layer of the data in the normalized scores within the PC-space. Peeling is an iterative process where the most external convex hull (first layer, $l = 1$) is removed and considering the remaining samples a new convex hull is computed (second layer, $l = 2$). The process is repeated, until there are not enough points left to continue. The remaining points, if present, are given a weight of 0. The samples belonging to each convex hull are inversely weighted with their respective peel number (weights equal to $1/l$). In $\boldsymbol{W}_c$, the samples of the first peel (ESP) have a weight 1 and for the last and inner most peel, a weight close to zero is set. The lower the relevance of the sample towards the MCR solution, the lower its weight.

2.4 Residual bootstrap analysis

Residual bootstrap analysis for regression [14] is a randomised error resampling technique to determine the stability of a model from a single data set ($\boldsymbol{D}$). The bootstrap framework is shown below, with Eq. 6-9:

The reconstructed data matrix $\widehat{\boldsymbol{D}}$ $(I \times J)$ is estimated from a singular value decomposition (SVD) [15] of $\boldsymbol{D}$ with k components.

(Eq. 6) $$\widehat{\boldsymbol{D}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{T}$$

Where the $\boldsymbol{U}$ $(I \times k)$ and $\boldsymbol{V}$ $(J \times k)$ are the left and right singular vectors of $\boldsymbol{D}$, respectively, and $\boldsymbol{\Sigma}$ $(k \times k)$ is a square matrix with its singular values on the diagonal and zeros elsewhere.

The error $\boldsymbol{E}$ $(I \times J)$ is calculated by subtracting $\widehat{\boldsymbol{D}}$ from $\boldsymbol{D}$.

(Eq. 7) $$\boldsymbol{E} = \boldsymbol{D} - \widehat{\boldsymbol{D}}$$

$\boldsymbol{E}$ is resampled to generate a new error matrix $\boldsymbol{E}_{bs}$ $(I \times J)$, by removing a random subset (1%) of samples from $\boldsymbol{E}$ and repopulating it with another random subset of $\boldsymbol{E}$. In this way a new error matrix is generated, following the same error distribution that is present within $\boldsymbol{E}$.

(Eq. 8) $$\boldsymbol{E} \rightarrow \boldsymbol{E}_{bs}$$

This resampled error is added back to $\widehat{\boldsymbol{D}}$ to generate a residual-bootstrapped data matrix $\boldsymbol{D}_{bs}$ $(I \times J)$ which is further processed or analysed, in this case, by means of MCR-ALS.

(Eq. 9) $$\boldsymbol{D}_{bs} = \widehat{\boldsymbol{D}} + \boldsymbol{E}_{bs}$$

To get a proper estimation of the model stability, the bootstrap is repeated 50 times [11], to generate 50 matrices $\boldsymbol{D}_{bs}$.

3 Data sets

Three data sets are analysed, two resulting from simulations and one from a six-component Raman hyperspectral image of a pharmaceutical tablet.

Data set 1

A set of three spectral profiles (figure 1a, 120 variables) and three concentration profiles (2595 samples) are simulated. The concentration profiles (equal for each component) span the entire mixture space, containing a set of pure, binary, and ternary samples. One pure sample per component is present and at least one spectral variable is fully selective. All three components have the exact same concentration distribution. The concentration and spectral profiles are multiplied with each other to obtain a noiseless data matrix. Afterwards, Gaussian noise (15 % of the signal intensity) is added to obtain a final data matrix (figure 1b). A clear triangle is observed within the normalised PC-scores plot (figure 1c), which indicates that every possible combination of the three components is present within the data matrix. 50 data matrices are generated, with each matrix having an error structure randomly sampled from a Gaussian distribution.

Data set 2

The simulated data matrix is generated as reported in Vitale et al. [4], which results in a three-component (A, B and C) system and features 56700 samples and 120 spectral-like variables (figure 2a). A set of 50 matrices are generated from it, by recalculating the error, but maintaining the exact same concentration and spectral profiles. The relative amount of noise is kept at 15% of the signal intensity, similarly to Data set 1. Component C is set as a minor component, meaning that a big portion of the samples contains mainly components A and B, and component C has a very low concentration across the samples. However, differently from the data matrices generated by Vitale et al., this data set contains 800 pure samples of C, this is because the noise level is three times larger. The spectra of a single data matrix are shown in figure 2b.

Data set 3

This data set relates to a six component semi-synthetic Raman image of a pharmaceutical tablet and consists of 5000 samples and 1600 variables. We refer to Coic et al. [8] for the details on the analysis. The spectra are pre-processed with a Savitzky-Golay filter [16], using a first order polynomial and a window size of 11. The six blended chemical compounds are known, and their corresponding spectral profiles (used as a reference) are taken from an in-house database. See figure 3 for an overview of the data. As can be seen in the normalized scores plot (figure 3c), the data structure shows that minor components are present, although a higher dimensional representation would be needed for a full visualisation.

A bootstrap analysis is performed on the data to obtain 50 bootstrapped matrices, as described in section 2.4 with $k = 10$.

4 Results and Discussion

For each data set, 50 MCR solutions (every model estimated from a matrix with a different error structure) are obtained by the approaches tested: 1) MCR-ALS on the data set (results denoted as "Full" in the remainder of the text); 2) weighted MCR-ALS with weights encoding ESP selection ("ESP") and 3) weighted MCR-ALS with weights encoding the results of convex peeling ("Weighted"). The dispersion of the solutions obtained from the 50 replicates can be compared among the different approaches to determine the stability of the estimated parameters.

For Data set 1, results are provided in figure 4. As expected, those obtained from "Full" show that without any weighting or selection, a good (accurately representing the ground-truth for each component) and stable (no dispersion) estimation of the pure spectral profiles is obtained. The solutions obtained from "ESP" show that weighting the ESPs as one and all others as zeros clearly has an impact on the dispersion of the solutions, indicating an increased variance in the estimates of the MCR-ALS model parameters (profiles). While for "Weighted", the performance is found to be similar to "Full".

In figure 5 the results for Data set 2 are shown, which now highlight, differently from Data set 1, the potential impact of an under-represented minor component on the accuracy of the outcomes [4]. The solutions obtained from "Full" show that MCR-ALS is not able to accurately estimate the parameters of component C, even though little to no dispersion is observed. With "ESP", the spectra of component C are estimated properly and point out the importance of selecting relevant samples to drive the MCR-ALS solutions towards the true one in the presence of minor components. However, this comes at the price of a higher dispersion, as already noted for Data set 1. For "Weighted", similarly to "ESP", accurate spectra are obtained, without any bias in the minor component estimate. However, in contrast to "ESP", very little dispersion in the model parameters is seen, since the full data set is used.

Figure 6 shows the results for Data set 3. Like in Coic et al. [8], "Full" cannot estimate all components, minor components C and F (which explain 0.05 and 1.77% of the variance of the original data, respectively) are missed. By contrast, "ESP" and "Weighted" can retrieve solutions very close to the reference spectra, with "Weighted" showing a decrease in the dispersion of the solutions compared to "ESP". These results corroborate the ones obtained

from Data set 2: a decrease in the dispersion of the parameter estimates of an order of magnitude for component F to around half for component B is observed. Only component C sees no decrease in dispersion, because "ESP" selects all the samples containing C, meaning that using the full data with respect to ESP adds no additional information on C. Concerning component F, "ESP" still selects the purest samples, however the selected samples have a significant noise level, inducing a dispersion in the solutions. Weighting the data set with convex peeling instead of just the ESPs, increases the number of analysed samples containing component F, in turn, reducing the variance in its calculated spectral profile.

When comparing the results of "Full" and "ESP", both Data sets 2 and 3 show that, in the presence of minor components, a trade-off is present between the approaches. One should choose between precise but biased solutions with "Full" or accurate but imprecise solutions with "ESP". "Weighted", instead, takes the middle ground, where the utilisation of the full data combined with the knowledge on the essential information they carry, in the presence of noise and minor components gives both more accurate and precise solutions.

5 Conclusion

With this work we show that, in the presence of minor components, ESP selection is required to drive the MCR-ALS solution towards the true one, with the caveat of losing parameter stability due to instrumental noise. We propose an extended weighting scheme within the weighted MCR-ALS framework that is based on convex hull data peeling and is able to preserve the benefits of ESP selection without reducing model parameter stability. This weighting framework is based on the relevance of the entire ensemble of investigated samples towards the MCR-ALS resolution. However, this can be further optimized by e.g., limiting the number of convex peels or applying a threshold on the sample relevance criterion to compress the data more adequately, reducing computation times. Furthermore, other relevance criteria can be applied as well (see the recent work done by Zade et al. [17]).

Acknowledgements

The authors acknowledge Laureen Coïc and Eric Ziemons for making their data available, and Nematollah Omidikia and Mathias Sawall for fruitful discussion. Raffaele Vitale and Cyril Ruckebusch acknowledge financial support from the 'ANR-21-CE29-0007' project (Agence Nationale de la Recherche).

References

[1] de Juan, A., & Tauler, R. (2021). Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review. Analytica Chimica Acta, 1145, 59–78. https://doi.org/10.1016/j.aca.2020.10.051

[2] Wentzell, P. D., Karakach, T. K., Roy, S., Martinez, M. J., Allen, C. P., & Werner-Washburne, M. (2006). Multivariate curve resolution of time course microarray data. BMC Bioinformatics, 7(1). https://doi.org/10.1186/1471-2105-7-343

[3] Blanchet, L., Réhault, J., Ruckebusch, C., Huvenne, J. P., Tauler, R., & de Juan, A. (2009). Chemometrics description of measurement error structure: Study of an ultrafast absorption

spectroscopy experiment. Analytica Chimica Acta, 642(1–2), 19–26. https://doi.org/10.1016/j.aca.2008.11.039

[4] Vitale, R., & Ruckebusch, C. (2022). On a black hole effect in bilinear curve resolution based on least squares. Journal of Chemometrics. https://doi.org/10.1002/cem.3442

[5] Ghaffari, M., Omidikia, N., & Ruckebusch, C. (2019). Essential Spectral Pixels for Multivariate Curve Resolution of Chemical Images. Analytical Chemistry, 91(17), 10943–10948. https://doi.org/10.1021/acs.analchem.9b02890

[6] Ruckebusch, C., Vitale, R., Ghaffari, M., Hugelier, S., & Omidikia, N. (2020). Perspective on essential information in multivariate curve resolution. TrAC Trends in Analytical Chemistry, 132, 116044. https://doi.org/10.1016/j.trac.2020.116044

[7] Sawall, M., Ruckebusch, C., Beese, M., Francke, R., Prudlik, A., & Neymeyr, K. (2022). An active constraint approach to identify essential spectral information in noisy data. Analytica Chimica Acta, 1233, 340448. https://doi.org/10.1016/j.aca.2022.340448

[8] Coic, L., Sacré, P. Y., Dispas, A., De Bleye, C., Fillet, M., Ruckebusch, C., Hubert, P., & Ziemons, R. (2022). Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations. Analytica Chimica Acta, 1198, 339532. https://doi.org/10.1016/j.aca.2022.339532

[9] Nardecchia, A., & Duponchel, L. (2020). Randomised SIMPLISMA: Using a dictionary of initial estimates for spectral unmixing in the framework of chemical imaging. Talanta, 217, 121024. https://doi.org/10.1016/j.talanta.2020.121024

[10] Ghaffari, M., Omidikia, N., & Ruckebusch, C. (2021). Joint selection of essential pixels and essential variables across hyperspectral images. Analytica Chimica Acta, 1141, 36–46. https://doi.org/10.1016/j.aca.2020.10.040

[11] González-Martínez, J. M., Camacho, J., & Ferrer, A. (2013). Bilinear modelling of batch processes. Part III: parameter stability. Journal of Chemometrics, 28(1), 10–27. https://doi.org/10.1002/cem.2562

[12] Preparata, F. P., & Shamos, M., I. (2012). Computational Geometry: An Introduction (Monographs in Computer Science) (Softcover reprint of the original 1st ed. 1985). Springer.

[13] Bro, Rasmus & Jong, Sijmen. (1997). A Fast Non-negativity-constrained Least Squares Algorithm. Journal of Chemometrics. 11. 393-401.

[14] Efron, B., & Tibshirani, R. J. (1993). An Introduction to the Bootstrap: Chapter 9 (Chapman & Hall/CRC Monographs on Statistics and Applied Probability) (1st ed.).

[15] Stewart, G. W. (1993). On the Early History of the Singular Value Decomposition. SIAM Review, 35(4), 551–566. https://doi.org/10.1137/1035134

[16] Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, 36(8), 1627–1639. https://doi.org/10.1021/ac60214a047

[17] Zade, S. V., Neymeyr, K., Sawall, M., Fischer, C., & Abdollahi, H. (2022). Data point importance: Information ranking in multivariate data. Journal of Chemometrics. https://doi.org/10.1002/cem.3453

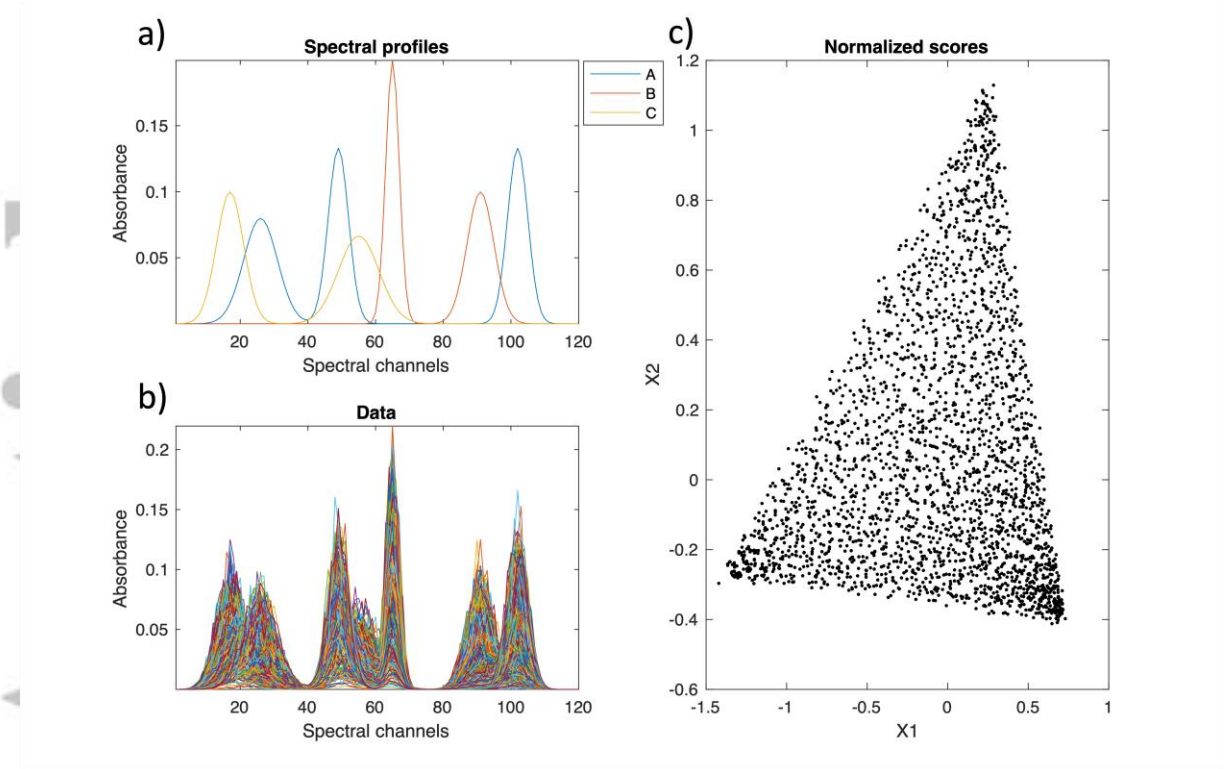

Figure 1

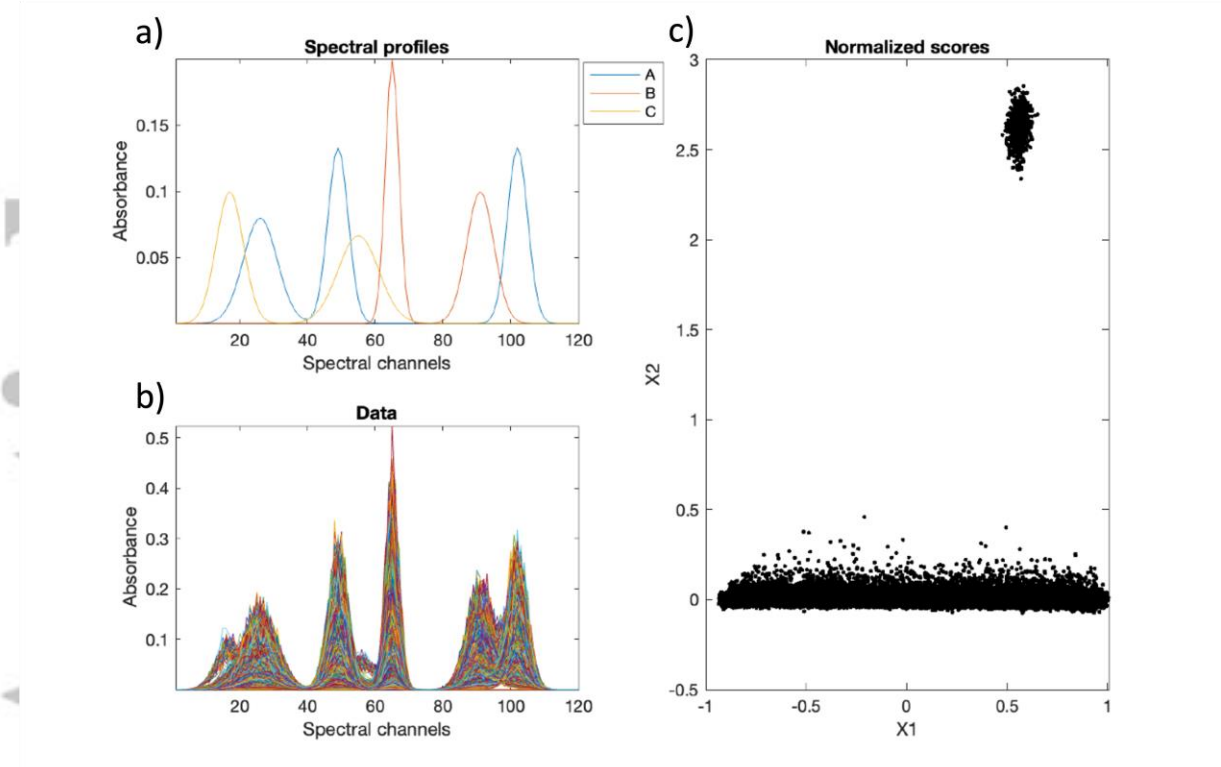

Figure 2

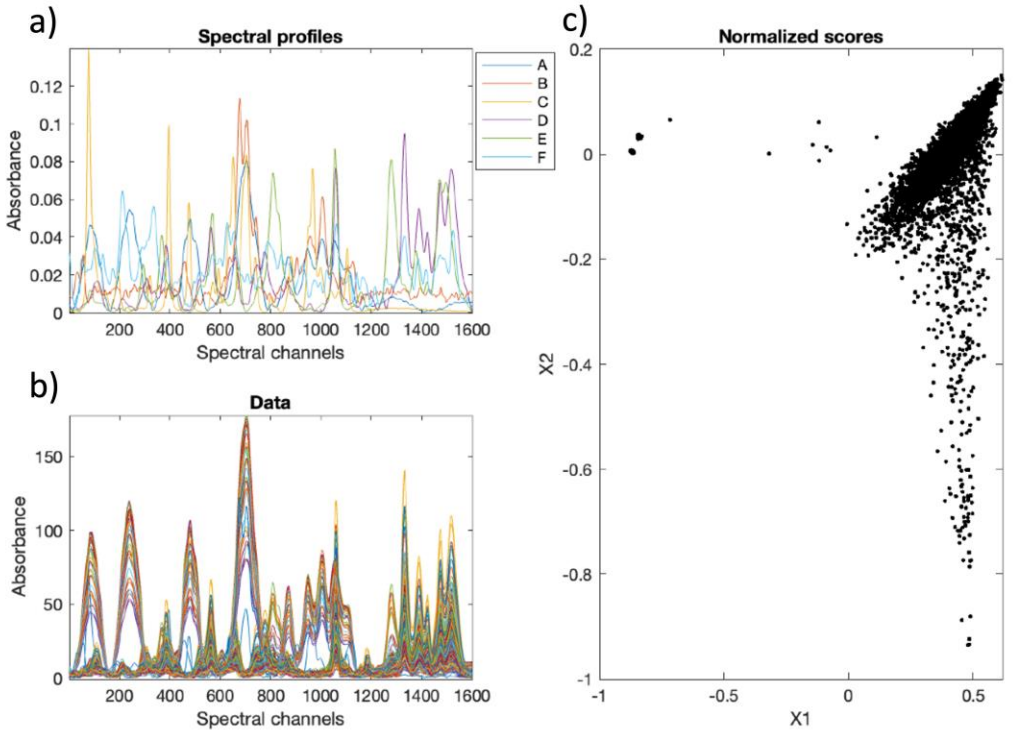

Figure 3

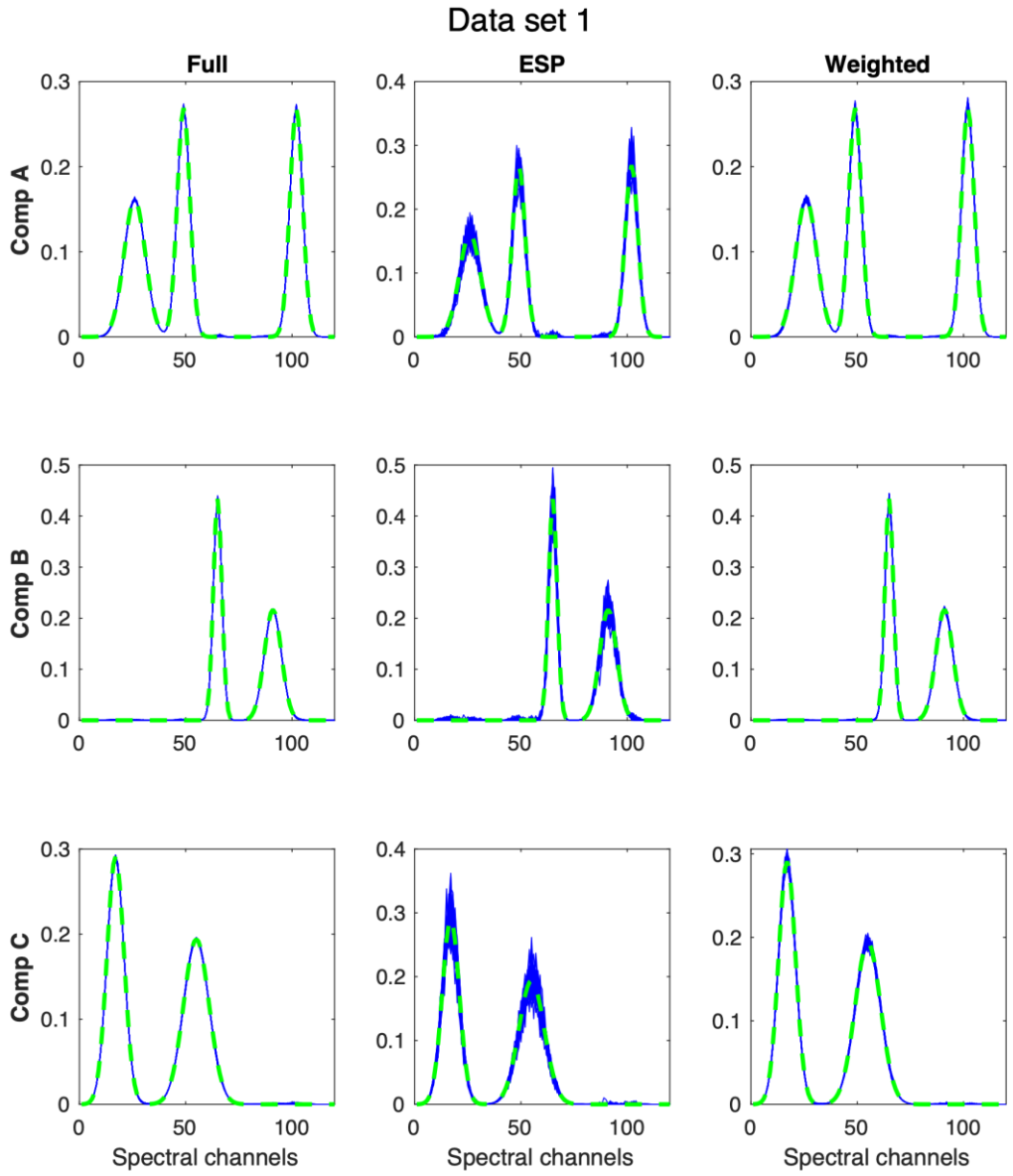

Figure 4

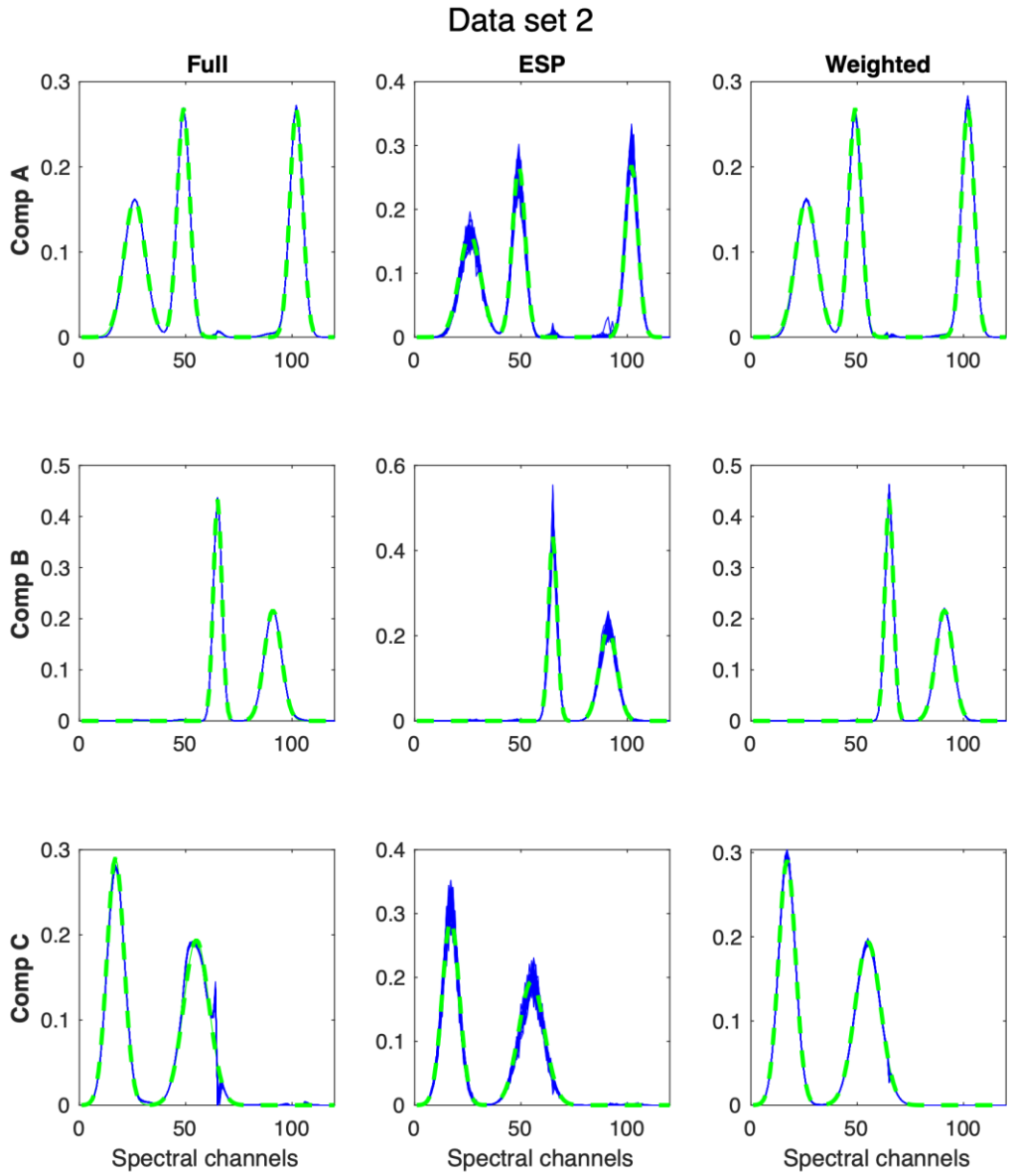

Figure 5

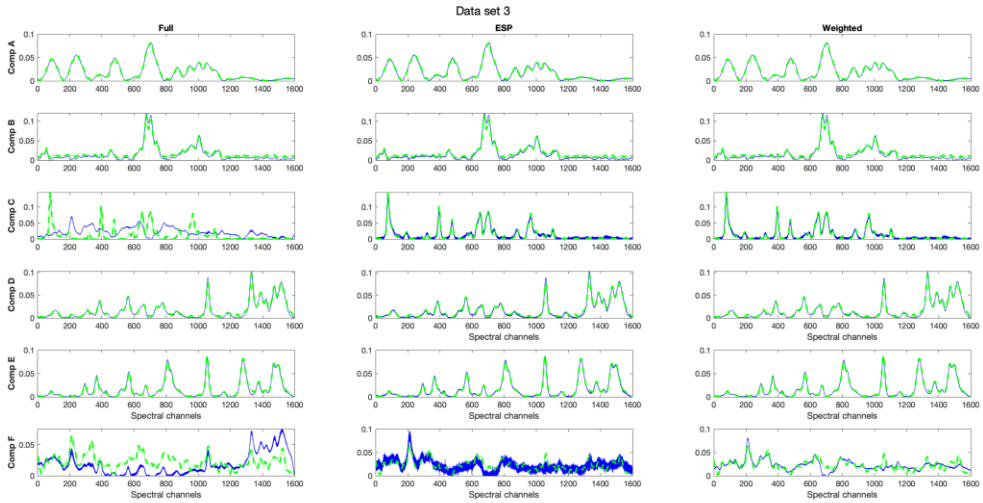

Figure 6