On a black hole effect in bilinear curve resolution based on least squares

Raffaele Vitale, Cyril Ruckebusch

To cite this version:

Raffaele Vitale, Cyril Ruckebusch. On a black hole effect in bilinear curve resolution based on least squares. Journal of Chemometrics, 2022, Journal of Chemometrics, -, 10.1002/cem.3442. hal-04511599

HAL Id: hal-04511599
https://hal.univ-lille.fr/hal-04511599

Submitted on 19 Mar 2024

Received: 22 April 2022 | Revised: 26 July 2022 | Accepted: 6 September 2022

DOI: 10.1002/cem.3442

SHORT COMMUNICATION

Journal of CHEMOMETRICS WILEY

On a *black hole* effect in bilinear curve resolution based on least squares

Raffaele Vitale | Cyril Ruckebusch

Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Lille, France

Correspondence
Raffaele Vitale, Univ. Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France.
Email: raffaele.vitale@univ-lille.fr

Funding information
The authors acknowledge funding form the project 'ANR-21-CE29-0007' (Agence Nationale de la Recherche).

Abstract

Least squares-based estimations lay behind most chemometric methodologies. Their properties, though, have been extensively studied mainly in the domain of regression, in relation to which the effect of well-known deleterious factors (like object leverage or data distributions deviating from ideal conditions) on the accuracy of the prediction of an external response variable has been thoroughly assessed. Conversely, much less attention has been paid to what these factors might yield in alternative scenarios, where least squares approaches are still utilised, yet the objectives of data modelling may be very different. As an example, one can think of multivariate curve resolution (MCR) problems which are usually addressed by means of multivariate curve resolution-alternating least squares (MCR-ALS). In this respect, this article wants to offer a perspective on the basic principles of MCR-ALS from the regression point of view. In particular, the following critical aspects will be highlighted: (i) in the presence of minor components, if the number of analysed data points is too large, the leverage of those that may be essential for a MCR-ALS resolution might become too low for guaranteeing its correctness, and (ii) in order to overcome this *black hole* effect and improve the accuracy of the MCR-ALS output, data *pruning* can be exploited. More in detail, this communication will provide a practical illustration of such aspects in the field of hyperspectral imaging where even single experimental runs may lead to the generation of massive amounts of spectral recordings.

KEYWORDS

curve resolution, least squares, leverage, regression

1 | INTRODUCTION: THE *BLACK HOLE* EFFECT

Several application studies recently reported in literature have highlighted how least squares-based methods for multivariate curve resolution (MCR), like multivariate curve resolution-alternating least squares (MCR-ALS[1,2]), might suffer

* Raffaele Vitale and Cyril Ruckebusch are equally contributing authors.

from critical limitations when coping with mixture datasets featuring so-called *minor* components, for example, chemical compounds observable only in correspondence of few pixels of a hyperspectral image.[3,4] In order to fully grasp the main reasons behind this particular issue, imagine a specimen composed by three different compounds (A, B and C) is actually to be characterised through a hyperspectral imaging experiment (see, e.g., Figure 1A,B). If the resulting (unfolded) hyperspectral data structure (say, **X**) is decomposed by principal component analysis (PCA[5,6]) as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

and the PCA scores obtained from Equation (1) are normalised so that all the columns of **T** are divided element-wise by the first column of **T**, denoted as $\mathbf{t}_1$:

$$\tilde{\mathbf{T}} = \mathbf{T} \oslash \mathbf{t}_1 \mathbf{1}^{\mathrm{T}} \tag{2}$$

with $\mathbf{1}^{\mathrm{T}}$ being a row vector of ones of appropriate dimensionality and $\oslash$ the element-wise (Hadamard) division operator, then the representation of the second- and third-component normalised scores enables an immediate and easy visualisation of the geometry of the specific MCR problem at hand.[7,8] In fact, under this normalisation constraint, provided that the whole set of image pixels spans all the possible mixture combinations of A, B and C, the second- versus third-component scores point cloud assumes a triangular (*simplex*) shape whose vertices actually correspond to the three pure-compound spectral pixels (labelled as 'A', 'B' and 'C', respectively; see Figure 1C). For any linear resolution approach, therefore, it would only be needed to somehow identify such vertices for accomplishing the spectral unmixing of **X**. For the sake of illustration, the blue solid lines in Figure 1C connote the solution yielded by MCR-ALS in this contingency (see also Figure 1D).[†]

Suppose now that the concentration of C gradually decreases all over the scanned surface, but that a single pure pixel for it still exists (see Figure 1E,H). This translates into peculiar distributions of the normalised projection scores yielded by the PCA decomposition in Equation (1) (see Figure 1F,I): the original simplex becomes, indeed, only partially covered, with scores more and more compacted along the direction connecting the pure A and pure B spectral pixels and a single scattered observation in the upper right part of the graph (the pure C spectral pixel).[‡] Here, the performance of MCR-ALS progressively worsens (see Figure 1F,G,I,J): in the scenario illustrated in Figure 1I, in spite of the fact that MCR-ALS is initialised with the spectral profiles of pure A, pure B and pure C, an accurate factorisation of the analysed data cannot even be attained—it has to be noticed that this can occur even in the absence of selective information for C (see supporting information). Such a phenomenon could also be interpreted in the following way: the increasing density of data points between 'A' and 'B' somehow *attracts* to a growing extent the MCR-ALS solutions towards the center of mass of the displayed data clouds, similarly to the effect of a black hole. The main reasons behind it are to be found in the least squares nature of the MCR-ALS algorithmic procedure.

2 | A LEVERAGE PROBLEM

If one were coping with a multivariate regression problem, looking at Figure 1I, the data point 'C' would appear as exhibiting a strong outlying behaviour. Nevertheless, given the high sample size, this outlying behaviour would not dramatically affect the results provided by the application of any least squares methodology.[10,11] In order to clarify this aspect, the concept of *leverage* needs to be introduced. The leverage of a given data item is commonly defined as the *n*-th diagonal element of the squared matrix **H**[12]:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \tag{3}$$

In general, it is said that the higher a leverage value, the more a least squares estimation based on **X** is influenced by the corresponding data point. **H** is characterised by a key property.[13–15] Its trace (i.e., the sum of all its diagonal elements), indeed, is always equal to the rank of **X**:

[†] In this article, MCR-ALS is always applied imposing only nonnegativity constraints.

[‡] Notice that the number of data points represented in Figure 1C,F,I is constant.

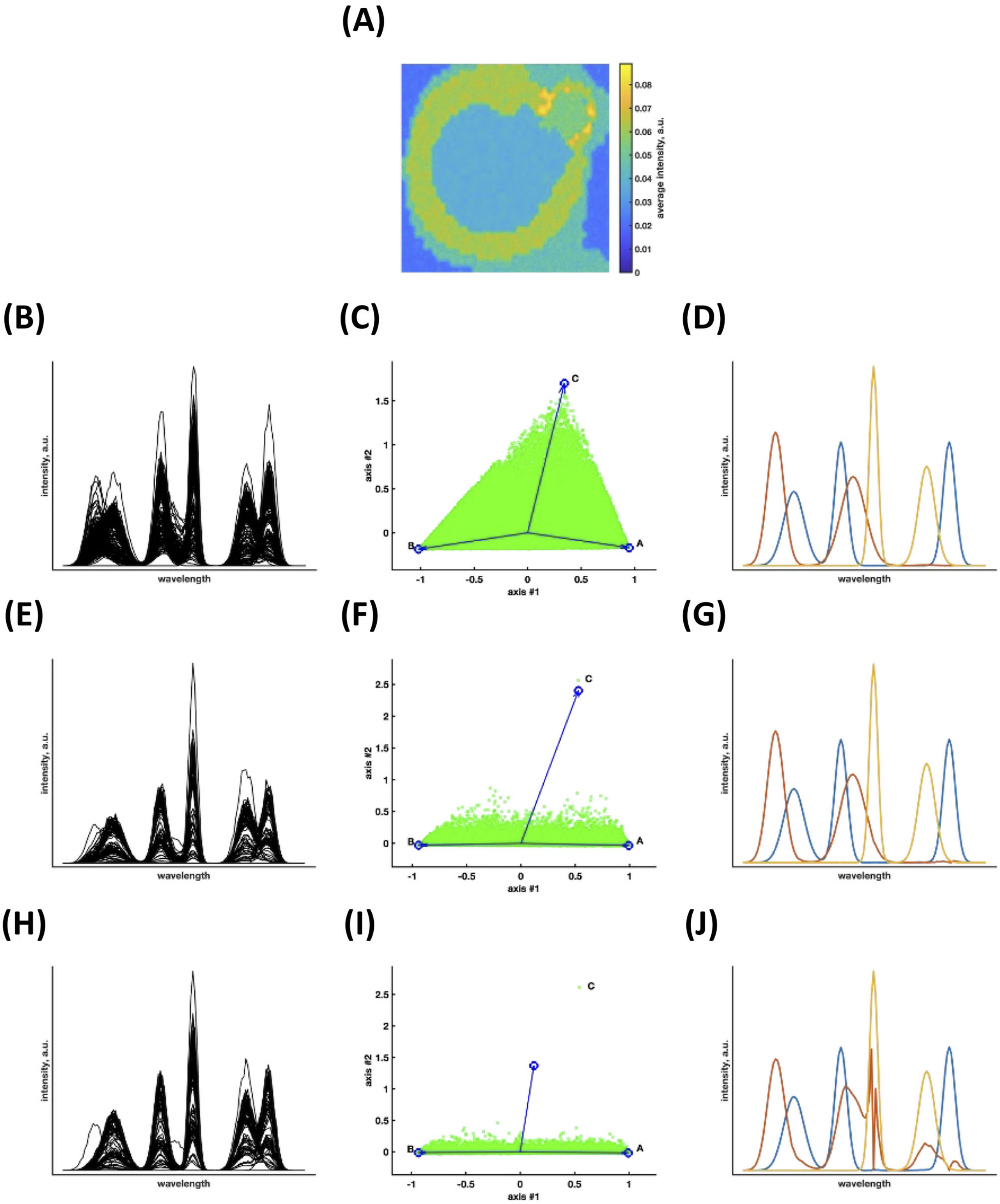

FIGURE 1 Legend on next page.

FIGURE 1 A simulated three-component hyperspectral imaging case-study: (A) illustrative wavelength-averaged hyperspectral image; (B) selection of pixel spectral profiles of a simulated image underlain by all possible mixture combinations of three compounds or ingredients (A, B and C); (C) normalised scores space representation of the resolved spectra (blue solid lines) returned by a multivariate curve resolution-alternating least squares (MCR-ALS) decomposition of the dataset Figure 1B relates to; (D) resolved spectra returned by a MCR-ALS decomposition of the dataset Figure 1B relates to; (E) selection of pixel spectral profiles of a simulated image generated accounting for a relatively low concentration of C all over the scanned scene and a single pure C spectral pixel; (F) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition of the dataset Figure 1E relates to; (G) resolved spectra returned by a MCR-ALS decomposition of the dataset Figure 1E relates to; (H) selection of pixel spectral profiles of a simulated image generated accounting for an extremely low concentration of C all over the scanned scene and a single pure C spectral pixel; (I) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition of the dataset Figure 1H relates to; (J) resolved spectra returned by a MCR-ALS decomposition of the dataset Figure 1H relates to. It is important to notice that, from a theoretical perspective,[9] (F) and (I) do not even reflect MCR solutions strictly fulfilling the nonnegativity constraint

$$\mathrm{tr}(\mathbf{H}) = \sum_{n=1}^{N} h_{n,n} = \mathrm{rank}(\mathbf{X}) \text{ with } 0 \leq h_{n,n} \leq 1 \tag{4}$$

with N being the number of observations in $\mathbf{X}$. Subsequently, if N grows while $\mathrm{rank}(\mathbf{X})$ is kept constant, $\mathrm{tr}(\mathbf{H})$ does not vary, but $\sum_{n=1}^{N} h_{n,n}$ involves a larger number of $h_{n,n}$ values. Therefore, on average, all $h_{n,n}$ decrease unless the leverage of the new objects included in $\mathbf{X}$ equals 0 (which barely happens in real case-studies). This property basically explains why increasing N (as per the common statement *the more the samples, the better the model*) is often exploited as a strategy to intrinsically reduce the bias that high-leverage data objects (if outliers) may generate and why, in the aforementioned situation, the influence of 'C' on the least squares procedure is limited. In an analogous way, here we propose to determine the leverage of a particular data point in the normalised scores subspace mentioned in Section 1 as the n-th diagonal element of the array $\mathbf{H}_{\tilde{\mathbf{T}}}$:

$$\mathbf{H}_{\tilde{\mathbf{T}}} = \tilde{\mathbf{T}}\left(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}}\right)^{-1}\tilde{\mathbf{T}}^{\mathrm{T}} \tag{5}$$

It is worth noticing that $\mathbf{H}$ and $\mathbf{H}_{\tilde{\mathbf{T}}}$ share the same mathematical features. In fact,

$$\mathrm{tr}\left(\mathbf{H}_{\tilde{\mathbf{T}}}\right) = \sum_{n=1}^{N} h_{\tilde{\mathbf{T}},n,n} = \mathrm{rank}\left(\tilde{\mathbf{T}}\right) \text{ with } 0 \leq h_{\tilde{\mathbf{T}},n,n} \leq 1 \tag{6}$$

and all $h_{\tilde{\mathbf{T}},n,n}$ usually decrease with N. Therefore, it is exactly for the same reason outlined before that, if N is too large, the leverage of data points that may be essential[16–18] in a MCR-ALS case-study (e.g., 'C') might become too low for a correct resolution to be achieved and that decreasing N could help improving the quality of the MCR-ALS output.

3 | A POSSIBLE WAY OUT: INFORMATION SELECTION

Given the properties of the diagonal elements of $\mathbf{H}$ and $\mathbf{H}_{\tilde{\mathbf{T}}}$, one potential strategy to somehow indirectly increase the leverage values of certain data points and, therefore, their importance in least squares-based algorithmic procedures is *pruning* the original set of measurements by reducing N while keeping $\mathrm{rank}(\mathbf{X})$ or $\mathrm{rank}\left(\tilde{\mathbf{T}}\right)$ unchanged. In the scenario illustrated in Figures 1I and 2B, for instance, if a relatively large portion of the observations in $\mathbf{X}$ (say, 65%) is randomly filtered out (ideally, without excluding the pure A, pure B and pure C spectral pixels; see Figure 2D), a significant improvement in the MCR-ALS solution could already be achieved (see Figure 2E,F). Nonetheless, in most cases, random pixel selection might be a suboptimal approach when MCR-ALS is to be run for hyperspectral image analysis.[3] A more adequate strategy to identify the most relevant spectral pixels for MCR relies on the estimation of the convex hull of the aforementioned normalised projection scores cloud.[19–21] When convex hull-based pruning is performed before the application of MCR-ALS to the data of Figure 2A, strikingly, the considerable reduction of the number of data points (see Figure 2G) leads to a correct and reliable unmixing of A, B and C (see Figure 2H,I). As also pointed out in

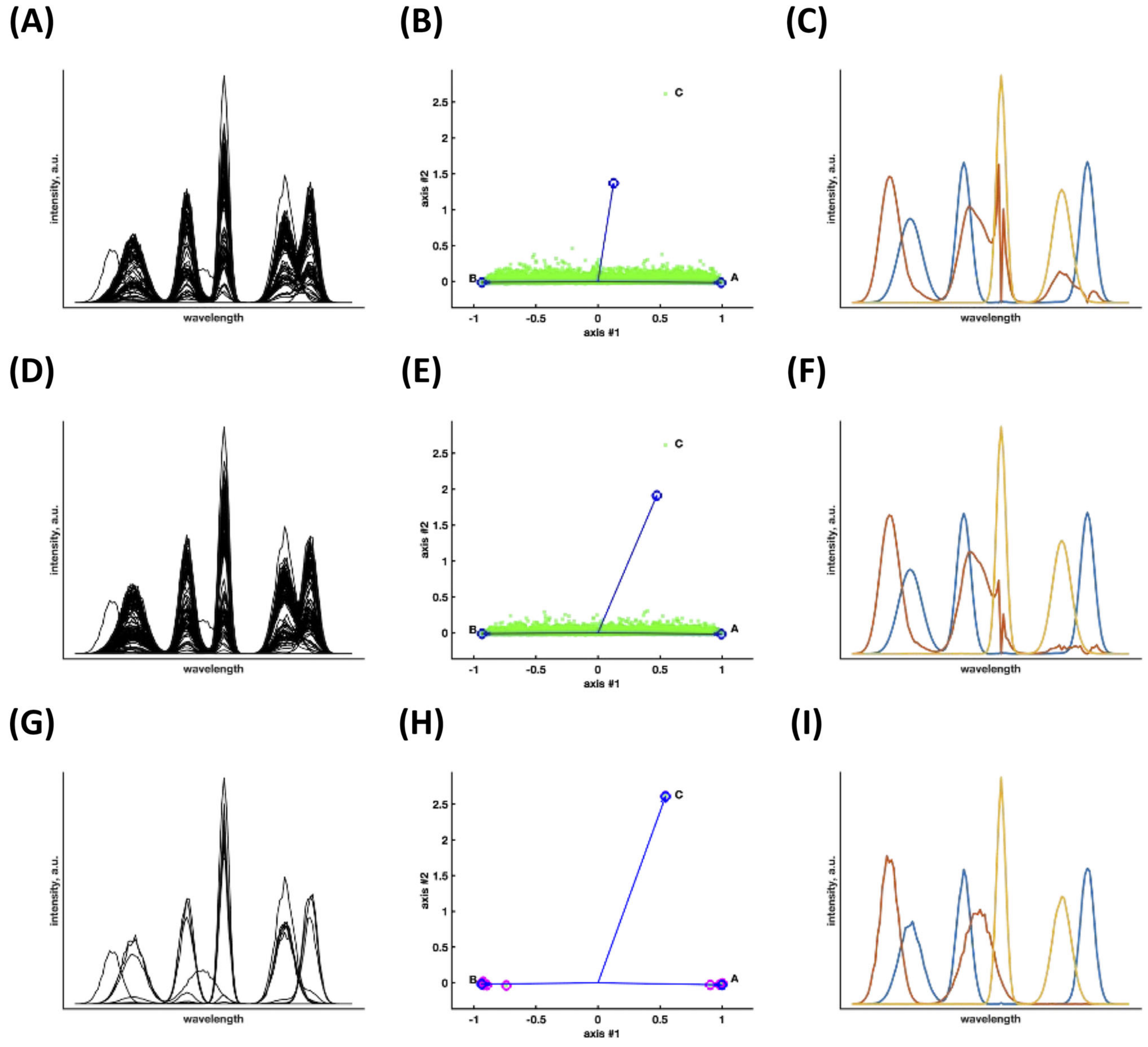

FIGURE 2 Performance of multivariate curve resolution-alternating least squares (MCR-ALS) before and after data pruning: (A) same as Figure 1H; (B) same as Figure 1I—the leverage of the observation labelled as 'C' is equal to 0.20 here ($N = 57,600$); (C) same as Figure 1J; (D) selection of spectral profiles drawn from a random subset of 35% of the original hyperspectral image pixels; (E) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition of the dataset Figure 2D relates to; the leverage of the observation labelled as 'C' is equal to 0.44 here ($N = 20,160$); (F) resolved spectra returned by a MCR-ALS decomposition of the dataset Figure 2D relates to; (G) spectral profiles of the most relevant pixels of the original hyperspectral image retrieved by the convex hull-based selection approach described in earlier studies[16–18]; (H) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition of the dataset Figure 2G relates to—the leverage of the observation labeled as 'C' is equal to 0.99 here ($N = 12$); (I) resolved spectra returned by a MCR-ALS decomposition of the dataset Figure 2G relates to. In (H), the data points actually considered for the MCR-ALS factorisation are highlighted by the magenta dots. For easing the comparison among Figure 2B,E,H, the same normalised scores subspace is preserved. It is important to notice that, from a theoretical perspective,[9] (B) and (E) do not even reflect MCR solutions strictly fulfilling the nonnegativity constraint

Section 1, a similar improvement can be observed even in the absence of pure C spectral profiles and irrespectively of the MCR-ALS initialisation strategy resorted to (see supporting information).[§]

We believe this to be the formal explanation of the effect observed in earlier studies.[3,4]

[§] It should be stressed that, in situations of higher measurement noise, it might be useful to retain not only the first but the first few convex layers of the normalised scores point cloud for a more robust estimation of the MCR-ALS solution from the reduced set of data.

4 | CONCLUSIONS

This featured communication was conceived in the attempt of clarifying an aspect (that, at a first glance, might seem counterintuitive) related to the effect that the number of analysed data points can have on the quality and the reliability of the solutions that least squares-based unmixing approaches may provide: in MCR scenarios, enhancing the importance of *extreme* measurement observations (increasing directly or indirectly their respective leverage values) can aid such approaches in achieving more accurate outcomes. Here, such an enhancement was accomplished by data pruning and information selection, but alternative strategies like object weighting based on measures of *essentiality* for the sake of curve resolution are currently being explored.[17,19,20] The implications of these strategies on the uncertainty and stability of the final results are, of course, of interest and will be investigated in future research.

ACKNOWLEDGEMENT

The authors acknowledge Laureen Coïc, Mahdiyeh Ghaffari, Mathias Sawall and Nematollah Omidikia for fruitful discussion.

DATA AVAILABILITY STATEMENT

Data simulation scripts are available upon request.

ORCID

Raffaele Vitale https://orcid.org/0000-0002-7497-1673
Cyril Ruckebusch https://orcid.org/0000-0001-8120-4133

REFERENCES

1. Tauler R, Smilde AK, Kowalski B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J Chemometr*. 1995;9:31-58.
2. De Juan A, Jaumot J, Tauler R. Multivariate curve resolution (MCR). Solving the mixture analysis problem. *Anal Methods*. 2014;6:4964-4976.
3. Coïc L, Sacré PY, Dispas A, et al. Pixel-based raman hyperspectral identification of complex pharmaceutical formulations. *Anal Chim Acta*. 2021;1155:338361.
4. Coïc L, Sacré PY, Dispas A, et al. Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations. *Anal Chim Acta*. 2022;1198:339532.
5. Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag*. 1901;2:559-572.
6. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933;24:417-441.
7. Grande BV, Manne R. Use of convexity for finding pure variables in two-way data from mixtures. *Chemometr Intell Lab*. 2000;50:19-33.
8. Rajkó R. Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method. *J Chemometr*. 2009;23:265-274.
9. Borgen OS, Kowalski B. An extension of the multivariate component-resolution method to three components. *Anal Chim Acta*. 1985;174:1-26.
10. Montgomery DC, Peck EA, Geoffrey Vining G. *Introduction to Linear Regression Analysis*. Hoboken, United States of America: John Wiley & Sons, Inc.; 2012.
11. Olivieri AC. *Introduction to Multivariate Calibration—A Practical Approach*. Cham, Switzerland: Springer Nature Switzerland AG; 2018.
12. Hoaglin DC, Welsch RE. The hat matrix in regression and ANOVA. *Am Stat*. 1978;32:17-22.
13. Gans P. *Data Fitting in the Chemical Sciences by the Method of Least Squares*. Chichester, United Kingdom: John Wiley & Sons, Ltd.; 1992.
14. Draper NR, Smith H. *Applied Regression Analysis*. Hoboken, United States of America: John Wiley & Sons, Inc.; 1998.
15. Freedman DA. *Statistical Models—Theory and Practice*. Cambridge, United Kingdom: Cambridge University Press; 2009.
16. Ghaffari M, Omidikia N, Ruckebusch C. Essential spectral pixels for multivariate curve resolution of chemical images. *Anal Chem*. 2019;91:10943-10948.
17. Ruckebusch C, Vitale R, Ghaffari M, Hugelier S, Omidikia N. Perspective on essential information in multivariate curve resolution. *Trend Anal Chem*. 2020;132:116044.
18. Ghaffari M, Omidikia N, Ruckebusch C. Joint selection of essential pixels and essential variables across hyperspectral images. *Anal Chim Acta*. 2021;1141:36-46.
19. Wentzell PD, Karakach TK, Roy S, Martinez MJ, Allen CP, Werner-Washburne M. Multivariate curve resolution of time course microarray data. *BMC Bioinformatics*. 2006;7:343.

20. Ahmad M, Vitale R, Cocchi M, Ruckebusch C. A solution based on sample weighting to the leverage problem in multivariate curve resolution-alternating least squares. XVIII Chemometrics in Analytical Chemistry, CAC 2022, 29/08-02/09/2022, Rome, Italy.
21. Sawall M, Kubis C, Schröder H, et al. Multivariate curve resolution methods and the design of experiments. *J Chemometr*. 2020; 34: e3159.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Vitale R, Ruckebusch C. On a *black hole* effect in bilinear curve resolution based on least squares. *Journal of Chemometrics*. 2023;37(2):e3442. doi:10.1002/cem.3442