FEATURED COMMUNICATION

On the *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, F-59000 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
None

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 have 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: in certain situations, i) 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* - *i.e.*, the reduction of the amount of observations of the investigated datasets - 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
leverage, least squares, regression, curve resolution

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 from critical limitations when coping with mixture datasets featuring so-called *minor* components, *e.g.* chemical

*Equally contributing authors.

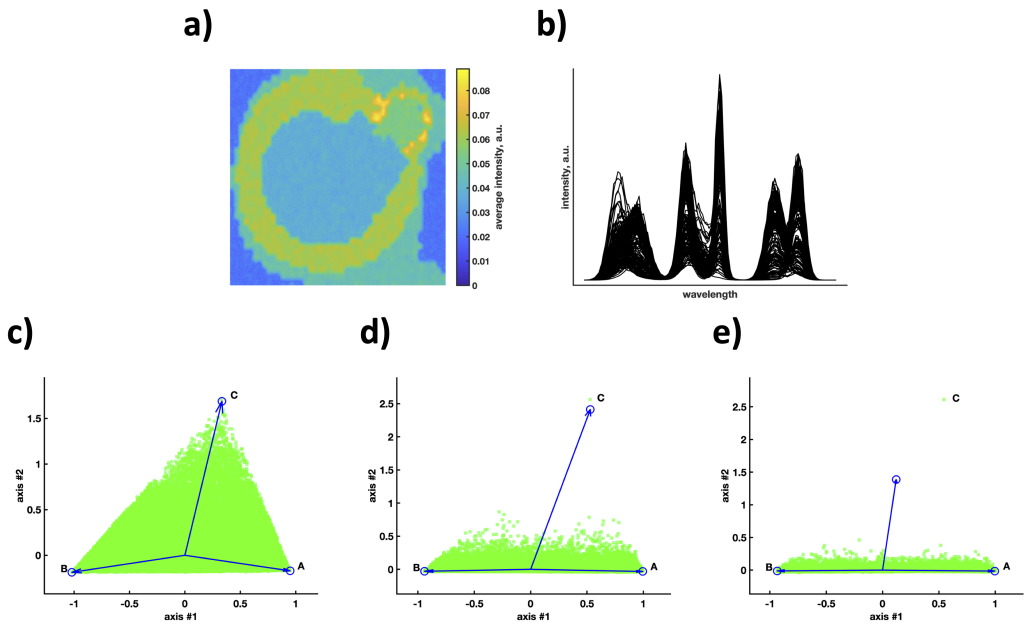

FIGURE 1 A simulated three-component hyperspectral imaging case-study: a) illustrative wavelength-averaged hyperspectral image; b) selection of pixel spectral profiles; c) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition of a simulated image whose pixels are underlain by all possible mixture combinations of three compounds or ingredients (A, B and C); d) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition 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; e) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition 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. It is important to notice that, from a theoretical perspective [5], d) and e) do not even reflect MCR solutions strictly fulfilling the non-negativity constraint.

compounds observable only in correspondence of few (pure) 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, for example, Figures 1a and 1b). If the resulting (unfolded) hyperspectral data structure (say, $\mathbf{X}$) is decomposed by Principal Component Analysis (PCA [6, 7]) as:

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

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

$$\tilde{\mathbf{T}} = \frac{\mathbf{T}}{\mathbf{t}_1 1^{\mathsf{T}}} \tag{2}$$

with 1^T being a row vector of ones of appropriate dimensionality, 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 [8, 9]. 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- *vs* 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 denote the solution yielded by MCR-ALS in this contingency[1].

Suppose now that the concentration of C gradually decreases all over the scanned surface, but that a single pure pixel for it still exists. This translates into peculiar distributions of the normalised projection scores yielded by the PCA decomposition in Equation 1 (see Figures 1d and 1e): 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)[2]. Here, the performance of MCR-ALS progressively worsens: in the scenario illustrated in Figure 1e, 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. This 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 1e, 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]:

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

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, 14, 15]. Its trace (*i.e.*, the sum of all its diagonal elements), indeed, always equals the rank of **X**:

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

with N being the number of observations in **X**. Subsequently, if N grows while rank (**X**) is kept constant, tr (**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 **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

[1] In this article, MCR-ALS is always applied imposing only non-negativity constraints.

[2] Notice that the number of data points represented in Figures 1c, 1d and 1e is constant.

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:

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

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 in a MCR-ALS case-study (for instance, "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 artificially 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 rank $(\mathbf{X})$ or rank $\left(\tilde{\mathbf{T}}\right)$ unchanged. In the scenario

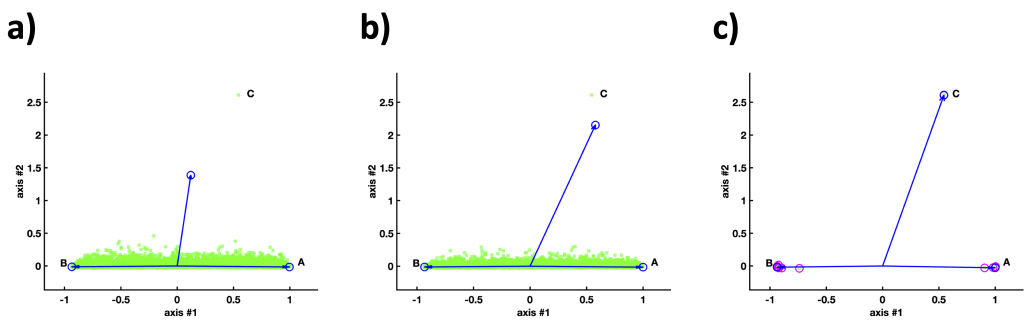

FIGURE 2 Performance of MCR-ALS before and after data pruning: a) same as Figure 1e - the leverage of the observation labelled as "C" equals 0.20 here (N = 57600); b) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition executed after the random selection of 35% of the image spectral pixels originally displayed in a) - the leverage of the observation labelled as "C" equals 0.44 here (N = 20160); c) normalised scores space representation of the resolved spectra (blue solid lines) returned by a MCR-ALS decomposition executed after the convex hull-based selection of the most relevant image spectral pixels originally displayed in a) - the leverage of the observation labelled as "C" equals 0.99 here (N = 12). In c), the data points actually considered for the MCR-ALS factorisation are highlighted by the magenta dots. For easing the comparison between the three plots, the same normalised scores subspace is preserved. It is important to notice that, from a theoretical perspective [5], a) and b) do not even reflect MCR solutions strictly fulfilling the non-negativity constraint.

illustrated in Figure 1e, for instance, if a relatively large portion of the observations in **X** (say, 65%) is randomly filtered out (ideally, without excluding the pure A, pure B and pure C spectral pixels), a significant improvement in the MCR-ALS solution could already be achieved (see Figure 2b). 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 [16, 17, 18]. When convex hull-based pruning is performed before the application of MCR-ALS to the data of Figure 1e, strikingly, the considerable reduction of the number of data points leads to a correct and reliable unmixing of A, B and C.
We believe this to be the formal explanation of the effect observed in the practical studies reported in [3, 4].

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 [19] could also be envisioned. 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.

Acknowledgements

The authors acknowledge funding form the project "ANR-21-CE29-0007" (Agence Nationale de la Recherche) as well as Laureen Coïc, Mahdiyeh Ghaffari and Nematollah Omidikia for fruitful discussion.

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, De Bleye C, Fillet M, Ruckebusch C, et al. Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations. Anal Chim Acta 2021;1155:article number 338361.

[4] Coïc L, Sacré PY, Dispas A, De Bleye C, Fillet M, Ruckebusch C, et al. Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations. Anal Chim Acta 2022;1198:article number 339532.

[5] Borgen OS, Kowalski B. An extension of the multivariate component-resolution method to three components. Anal Chim Acta 1985;174:1–26.

[6] Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag 1901;2:559–572.

[7] Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol 1933;24:417–441.

[8] Grande BV, Manne R. Use of convexity for finding pure variables in two-way data from mixtures. Chemometr Intell Lab 2000;50:19–33.

[9] Rajkó R. Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method. J Chemometr 2009;23:265–274.

[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:article number 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:article number 343.