A *black hole* effect in bilinear curve resolution based on alternating least squares

R. Vitale[1] L. Coic[1] C. Ruckebusch[1]
[1] Dynamics, Nanoscopy and Chemometrics (DyNaChem), LASIRE, CNRS, Univ. Lille, raffaele.vitale@univ-lille.fr

Keywords: leverage, least squares, regression, curve resolution.

1 Introduction

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 [1]. 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 [2]). In this respect, this work 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 reduction – *i.e.*, the selection of a smaller subset of observations among all the investigated ones – 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.

2 Motivation example

Imagine a situation in which a univariate regression model is to be constructed between an explanatory variable (x) and a response variable (y), both measured for 100 samples (see Figure 1a). In the presence of a single high-leverage data item (the empty green dot), the estimation of the parameters of this model results to be clearly biased if carried out by classical least squares (see the blue dotted line). Assume now to increase the number of collected data points up to 10000 (see Figure 1b). In this case, the bias induced by the aforementioned high-leverage observation becomes insignificant and, subsequently, the final regression model is capable of accurately describing the underlying correlation between x and y. *Nothing new under the sun*, one may say, but in a MCR context, the consequences of such a property could be dramatic, especially when analysing data related to chemical mixtures containing minor constituents. An example is provided in Figure 1c. Figure 1c displays the (first-eigenvector normalised) scores yielded by the Principal Component Analysis (PCA) factorisation of a set of spectral measurements conducted on mixtures of three ingredients, A, B and C, in which C always appears in very low concentrations except for a single sample that consists only of this component (see the green square labelled as “C”). Measurements of pure A and pure B are also available (see the green squares labelled as “A” and “B”, respectively). If a MCR-ALS decomposition of these data is carried out, the high density of points noticeable in the bottom area of the plot prevents the computational procedure from attaining the correct resolution

(see the blue arrows) even if selective information exists for A, B and C. One may look at this phenomenon as if such points would "attract" the MCR-ALS solution towards the centre of mass of the represented point cloud (hence the name *black hole effect*). It goes without saying that increasing even more the number of observations would further emphasize this effect. Conversely, if data are preliminarily reduced (here, only 9 data points were selected as proposed in [3]), thus, increasing the leverage value of the observation labelled as "C", more accurate and reliable resolved profiles are obtained by the execution of MCR-ALS (see Figure 1d).

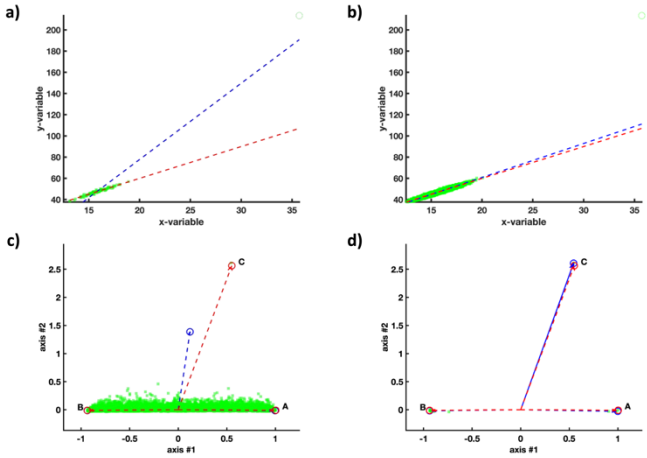

Figure 1 – Top panel: univariate regression plots in the presence of a single outlying observation. a) sample size: 100 – one high-leverage outlying data point; b) sample size: 10000 – one low-leverage outlying data point. Bottom panel: PCA representation of a multivariate curve resolution problem for ternary mixtures. c) sample size: 57600 – biased estimation of component C; d) sample size: 9 (numbered data points) – reliable estimation of component C. Original data are graphed as green symbols, ground-truth solutions as red dots, lines and arrows and estimated solutions (yielded by classical least squares and MCR-ALS, respectively) as blue dots, lines and arrows.

3 Results and discussion

Different MCR scenarios will be illustrated through the analysis of hyperspectral imaging datasets. It will be highlighted how an appropriate pixel selection (conducted prior to the modelling stage) can be crucial to attain reliable MCR-ALS resolutions, even in the presence of pure selective information encoded within the original measurements.

4 Conclusion

This presentation will offer a novel perspective on the MCR problem from the point of view of regression theory. Such a perspective will shed light on the effect that the size of the analysed data and their intrinsic multivariate distribution can have on the quality and the reliability of the solutions that least squares-based curve resolution approaches may provide. This will potentially open new interesting outlooks on an aspect not yet well-established in the chemometric community: information selection in MCR.

5 References

[1] H. Martens, T. Næs: *Multivariate Calibration*. John Wiley & Sons, Ltd., New York, United States of America, 1st Edition, 1989.

[2] Tauler, R., Smilde, A. & Kowalski, B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemom*. 9, 31-58, 1995.

[3] Ruckebusch, C., Vitale, R., Ghaffari, M., Hugelier, S. & Omidikia, N. Perspective on essential information in multivariate curve resolution. *TrAC Trends Anal. Chem*. 132, article number 116044, 2020.