Divide et impera: how disentangling common and distinctive variability in multi-set data analysis can aid industrial process troubleshooting and understanding

Raffaele Vitale[a,b,c,*], Onno E. de Noord[d,e], Johan A. Westerhuis[f], Age K. Smilde[f], Alberto Ferrer[a]

[a]*Grupo de Ingeniería Estadística Multivariante, Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain*
[b]*Molecular Imaging and Photonics Unit, Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001, Leuven, Belgium*
[c]*Laboratoire de Spectrochimie Infrarouge et Raman - UMR 8516, Université de Lille - Sciences et Technologies, Bâtiment C5, 59655, Villeneuve d'Ascq, France*
[d]*Shell Global Solutions International B.V., Shell Technology Centre Amsterdam, PO Box 38000, 1030 BN, Amsterdam, The Netherlands*
[e]*Advanced Data Analysis Consultancy, 2013 AW - 112, Haarlem, The Netherlands*
[f]*Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands*

Abstract

The possibility of addressing the problem of process troubleshooting and understanding by means of modelling common and distinctive sources of variation (*factors* or *components*) underlying two sets of measurements was explored in a real world industrial case-study. The used strategy includes a novel approach to systematically detect the number of common and distinctive components. An extension of this strategy for the analysis of a larger number of data blocks, which allows the comparison of data in multiple processing units, is also discussed.

Keywords: common components, distinctive components, permutation testing, Singular Value Decomposition (SVD), Canonical Correlation Analysis (CCA)

1. Introduction

Nowadays, industrial processes generate massive amounts of data, which are collected for on-line treatment or posterior analysis. In order to guarantee and preserve the high quality of the final

*Corresponding author:
Telephone number: +33769476654
Email address: rvitale86@gmail.com (Raffaele Vitale)

products and to minimise the number of failures, most manufacturing companies design monitoring schemes which allow abnormal events to be quickly, easily and efficiently recognised (*fault detection*) and their possible root causes to be correctly identified (*fault diagnosis*). After that, *ad hoc* countermeasures can be adopted to recover Normal Operating Conditions (NOC). These monitoring schemes are usually constructed through empirical approaches based on e.g., Principal Component Analysis (PCA) or Partial Least Squares regression (PLS) [1–5]. More specifically, i) data collected under NOC are used to calibrate a so-called *in-control* model, and afterwards ii) incoming data are projected onto its space for the assessment of the future evolution of the process. Once an *out-of-control* signal is spotted, it is fundamental to verify which of the measured variables are mostly affected by the fault. Tools like the so-called contribution plots [4] can be exploited for this purpose.

However, process understanding and troubleshooting can also be regarded from a slightly different perspective. Imagine that the same engineering variables (i.e., temperatures, pressures, flow rates, etc.) are resorted to for characterising the same industrial process i) during NOC and ii) during the occurrence of a failure. Subsequently, the two different data blocks resulting from the two distinct time periods could be *fused* and analysed as a multi-set structure. In particular, assuming that the variation that is distinctive for the second dataset contains information on a possible deviation from NOC, it may be possible to retrieve and explore such variation to find out what is causing the failure in production.

Distinguishing the common and distinctive sources of variability (*factors* or *components*) underlying, for instance, two sets of data has recently become an intriguing and challenging task [6, 7]. In the last decades, many dimensionality reduction methods have been proposed to model common and distinctive components when dealing with multi-set data analysis problems. Table 1 lists some of the most commonly used of these approaches, recently compared in [8]. These techniques can be classified according to their capability of handling various types of data structures (i.e., object-wise or variable-wise linked [8, 19]) and retrieving the distinctive components affecting the variability of the considered data blocks. Furthermore, most of them do not encompass preliminary computational steps aimed at identifying the number of such common and distinctive components, which in many cases can jeopardise the stability, and therefore the interpretation of

Table 1: List of some of the most commonly used dimensionality reduction methods for common and distinctive component modelling. The techniques are differentiated according to their capability of handling object-wise or variable-wise linked data structures and retrieving distinctive components affecting the variability of the data blocks under study. SCA, DISCO-SCA, GSVD, CCA, O2PLS, and JIVE stand for Simultaneous Component Analysis, DIStinctive and COmmon Simultaneous Component Analysis, Generalised Singular Value Decomposition, Canonical Correlation Analysis, 2-block Orthogonal Projections to Latent Structures, and Joint and Individual Variation Explained, respectively

	SCA [9, 10]	DISCO-SCA [11, 12]	Adapted GSVD [13, 14]	ECO-POWER [15]	CCA [16]	O2PLS [17]	JIVE [18]
Common components	✓	✓	✓	✓	✓	✓	✓
Distinctive components	✗	✓	✓	✗	✗	✓	✓
Object-wise linked data	✓	✓	✓	✓	✓	✓	✓
Variable-wise linked data	✓	✓	✓	✗	✗	✗	✗

the final results.

In the present article, the possibility of addressing the problem of process troubleshooting and understanding by means of modelling common and distinctive sources of variation will be explored in a real world industrial chemical case-study. A novel strategy to systematically detect the number of common and distinctive components when two blocks of measurements are dealt with will also be described.

2. Dataset

21 engineering variables (mainly temperatures, pressures and flow rates) were recorded over time in a single reacting unit during the evolution of 77 batches of a 6-stage process. The data were first synchronised by a recently proposed algorithm, Multisynchro [20, 21], to guarantee all these batches had the same evolution pace, and afterwards unfolded batch-wise [5]. The final two-way array was thereafter split into two different blocks, namely $\mathbf{X}_1$ (of dimensions $N_{\mathbf{X}_1} \times J$ where $N_{\mathbf{X}_1} = 20$ and $J = 9016$) and $\mathbf{X}_2$ ($N_{\mathbf{X}_2} \times J$ where $N_{\mathbf{X}_2} = 57$), having the same number of columns and whose single rows carry the whole time evolution of all the aforementioned variables for every process run. $\mathbf{X}_1$ contained data associated to batches that were manufactured during a first time period in which product quality was excellent and stable. The data in $\mathbf{X}_2$ were instead collected during a second time period when product quality fluctuated and gradually became worse. $\mathbf{X}_1$ and

$\mathbf{X}_2$ were then auto-scaled and scaled to equal sum-of-squares (i.e., block-scaled [10, 22]).
A similar data structure was available for a second set of batch runs of the same process manufactured in another reacting unit during the same time periods and for which the same engineering variables were monitored. Here, the size of the 2 different arrays, $\mathbf{Z}_1$ and $\mathbf{Z}_2$, was 22×9016 and 14×9016, respectively.

3. Results and discussion

3.1. Common and distinctive component modelling strategy

Suppose the information associated to the common and distinctive components of $\mathbf{X}_1$ and $\mathbf{X}_2$, has to be recovered. In a certain sense, considering the specific nature of the case-study at hand, this might be looked at as a supervised Multivariate Statistical Process Control (MSPC) problem in which both *in-control* and *out-of-control* sources of variation are to be captured at the same time to achieve better insights into the root causes of the process deviations experienced in the second manufacturing time period. Furthermore, apart from evaluating the nature of the common and distinctive components for each manufacturing unit separately, it could also be interesting to assess whether the distinctive factors from the second time period related to each unit have something in common, which would indicate that both units were affected by the same type of deviation. Such an aspect is very relevant in the present scenario, because the two reactors share the same run-down tank from which product is taken for the performance test. Figure 1 displays a general sketch of the data combination strategy presented in this paper. But how can one focus on the various sources of variability highlighted in this scheme? One possible way is the following:

1. the A_c common factors between $\mathbf{X}_1$ and $\mathbf{X}_2$ are modelled by applying SVD to $\mathbf{X}_1\mathbf{X}_2^T$:

$$\mathbf{X}_1\mathbf{X}_2^T = \boldsymbol{\Upsilon}_c\boldsymbol{\Sigma}_c\boldsymbol{\Psi}_c^T + \mathbf{E}_c \tag{1}$$

and projecting $\mathbf{X}_1$ and $\mathbf{X}_2$ onto $\boldsymbol{\Upsilon}_c$ and $\boldsymbol{\Psi}_c$, respectively, as:

$$\mathbf{V}_{1,c} = \mathbf{X}_1^T\boldsymbol{\Upsilon}_c \tag{2}$$

$$\mathbf{V}_{2,c} = \mathbf{X}_2^T\boldsymbol{\Psi}_c \tag{3}$$

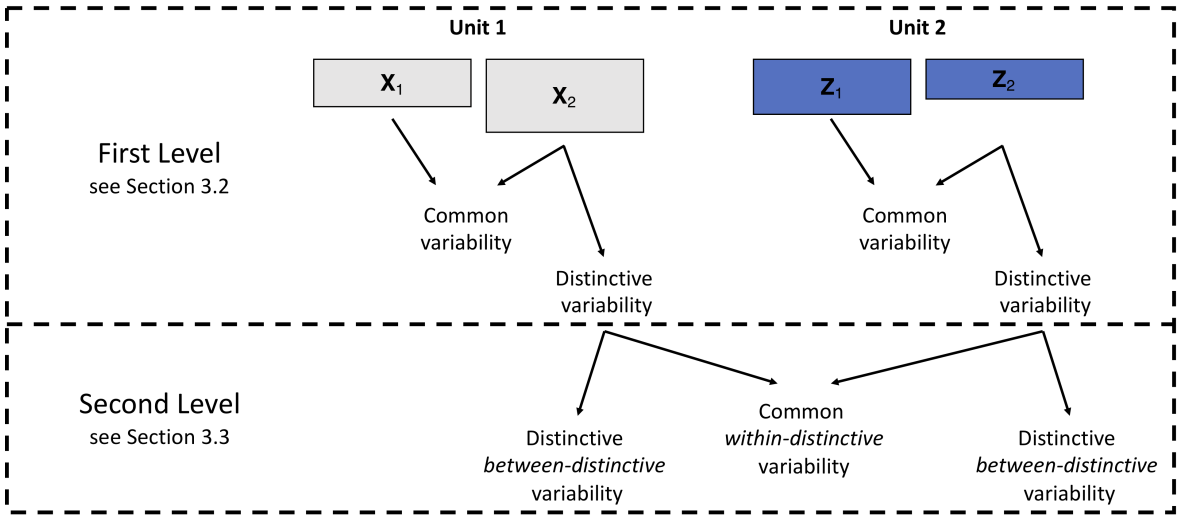

Figure 1: Schematic representation of the proposed data combination strategy. For the sake of clarity, the $\mathbf{X}_1/\mathbf{X}_2$ and $\mathbf{Z}_1/\mathbf{Z}_2$ matrices contain the evolution of the batches manufactured in the two different reacting units during the first and the second production period, respectively

with $\boldsymbol{\Upsilon}_c$ of dimensions $N_{\mathbf{X}_1} \times A_c$, $\boldsymbol{\Sigma}_c$ of dimensions $A_c \times A_c$, $\boldsymbol{\Psi}_c$ of dimensions $N_{\mathbf{X}_2} \times A_c$, $\mathbf{E}_c$ of dimensions $N_{\mathbf{X}_1} \times N_{\mathbf{X}_2}$ and $\mathbf{V}_{1,c}$ and $\mathbf{V}_{2,c}$ of dimensions $J \times A_c$;

2. the common components are deflated from $\mathbf{X}_1$ and $\mathbf{X}_2$ as:

$$\mathbf{X}_{1,d}^T = \mathbf{X}_1^T - \mathbf{X}_1^T \boldsymbol{\Upsilon}_c \boldsymbol{\Upsilon}_c^T \tag{4}$$

$$\mathbf{X}_{2,d}^T = \mathbf{X}_2^T - \mathbf{X}_2^T \boldsymbol{\Psi}_c \boldsymbol{\Psi}_c^T \tag{5}$$

3. SVD is finally used to retrieve the distinctive factors of $\mathbf{X}_1$ and $\mathbf{X}_2$:

$$\mathbf{X}_{1,d} = \mathbf{U}_{1,d} \mathbf{S}_{1,d} \mathbf{V}_{1,d}^T + \mathbf{E}_{1,d} \tag{6}$$

$$\mathbf{X}_{2,d} = \mathbf{U}_{2,d} \mathbf{S}_{2,d} \mathbf{V}_{2,d}^T + \mathbf{E}_{2,d} \tag{7}$$

where $\mathbf{U}_{1,d}$ is $N_{\mathbf{X}_1} \times A_{1,d}$-sized, $\mathbf{S}_{1,d}$ is $A_{1,d} \times A_{1,d}$-sized, $\mathbf{V}_{1,d}$ is $J \times A_{1,d}$-sized, $\mathbf{E}_{1,d}$ is $N_{\mathbf{X}_1} \times J$-sized, $\mathbf{U}_{2,d}$ is $N_{\mathbf{X}_2} \times A_{2,d}$-sized, $\mathbf{S}_{2,d}$ is $A_{2,d} \times A_{2,d}$-sized, $\mathbf{V}_{2,d}$ is $J \times A_{2,d}$-sized, and $\mathbf{E}_{2,d}$ is $N_{\mathbf{X}_2} \times J$-sized. $A_{1,d}$ and $A_{2,d}$ are estimated by permutation testing (see Appendix B) [23].

The procedure can be afterwards iterated for each level of the hierarchical structure depicted in Figure 1.

For attaining a reasonable guess of A_c, one can proceed as follows:

1. the total number of factors underlying $\mathbf{X}_1$ and $\mathbf{X}_2$ (A_1 and A_2) is calculated as for $A_{1,\mathrm{d}}$ and $A_{2,\mathrm{d}}$;
2. $\mathbf{X}_1$ and $\mathbf{X}_2$ are decomposed by SVD and their first A_1 and A_2 right singular vectors are retained, respectively;
3. the right singular vector matrices ($\mathbf{V}_1$, $J \times A_1$, and $\mathbf{V}_2$, $J \times A_2$) are then subjected to Canonical Correlation Analysis (CCA, see Appendix A) [16] (being the J-dimensional mode the shared one between $\mathbf{V}_1$ and $\mathbf{V}_2$[i]). The statistical significance of the resulting canonical correlations is evaluated through a permutation test carried out randomising iteratively the order of the entire rows of either $\mathbf{V}_1$ or $\mathbf{V}_2$ and recomputing the CCA solution. Canonical correlations larger than the 99$^{\mathrm{th}}$ percentile of their null-distributions are considered statistically significant. The number of statistically significant canonical correlations is set as A_{c}.

It has to be noticed that $\mathbf{V}_{1,\mathrm{c}}$ and $\mathbf{V}_{2,\mathrm{c}}$ do not exactly correspond to the canonical variates whose statistical significance is assessed. Nevertheless, assuming that common components show a relatively high correlation between blocks (either positive or negative), CCA can be utilised to get an at least tentative idea of their number before the proper data modelling phase, which can, in principle, be addressed by any of the methodologies mentioned in Section 1.

3.2. Unit 1 data analysis (first level)

The novel strategy was first applied to the process data collected in the first reacting unit. As described above, the dataset contains data on batch runs from two different time periods. In the second time period the product was still on specification, but in one particular performance test the quality started to fluctuate and gradually became worse. This performance test was not carried out on product from individual batch runs, but on blends from a run-down tank. Therefore, the question was formulated as: what has changed during the second time period in comparison to the first one, i.e., what is distinctive in this second time period? Assuming, as specified before, that the distinctive variation of the batches from the second time period contains information on the product quality issue, the idea was to i) determine the number of common components shared

[i]This can be considered a *trick* for adapting CCA when variable-wise linked data are coped with.

by the two data blocks and supposedly accounting for the normal variability of the process, ii) filter them from the second dataset by deflation and iii) explore the remainder trying to unveil possible causes of the deviation. Two and five factors were detected as statistically significant by the permutation-based effective rank estimation algorithm in the two datasets, respectively (i.e., the second time period was less homogeneous), and the presence of a single common component was found by the CCA-based permutation test. Since the attention is focused on the distinctive variability of the second time period, Figure 2 shows the time evolution of the loadings of its first distinctive component (found to be statistically significant after executing again the afore-mentioned effective rank estimation algorithm) for the 21 measured variables (first column of $\mathbf{V}_{2,\mathrm{d}}$

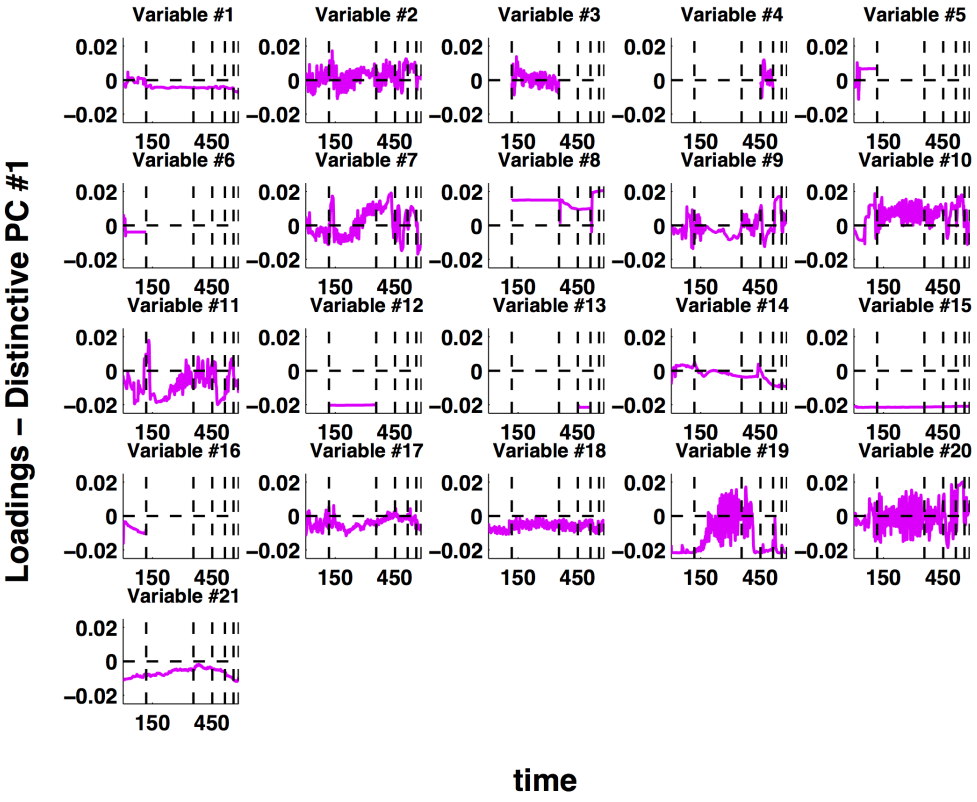

Figure 2: Industrial batch process data - Reacting unit 1: time profiles of the loadings of the first distinctive component of the second time period batch data block for the 21 measured variables (first column of $\mathbf{V}_{2,\mathrm{d}}$ according to Section 3.1). The vertical dashed lines separate the 6 stages of the industrial process. As not all the variables were active in these stages, part of such profiles is missing

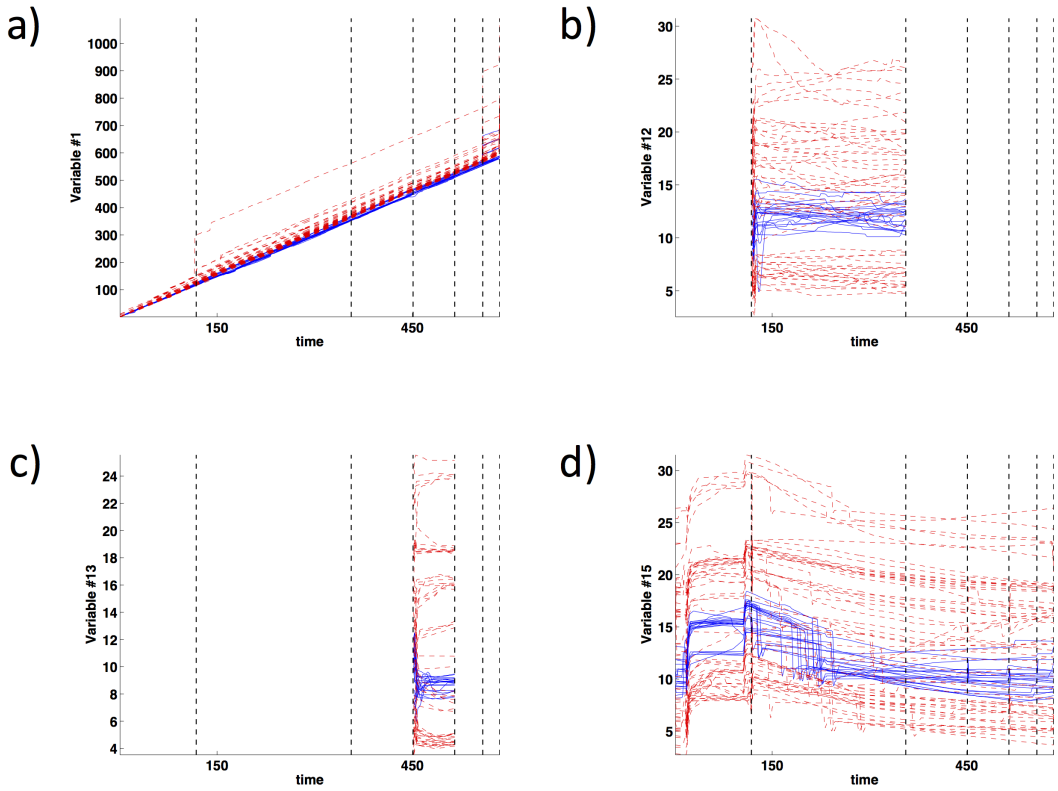

Figure 3: Industrial batch process data - Reacting unit 1: original time trajectories of variables a) #1, b) #12, c) #13 and d) #15 for the first period (blue solid lines) and the second period (red dashed lines) runs. The vertical dashed lines separate the 6 stages of the industrial process. As these variables were not active in every stage, part of their time trajectories is missing

according to Section 3.1)[ii]. Among those presenting a consistent non-zero temporal trend and thus a consistent contribution to this component, variables #1, #12, #13 and #15 generally exhibited both a higher variability and a higher average level in the second period batch runs than in the runs from the first time period (see Figure 3) and were isolated as those of interest from an engineering point of view. MSPC charts with contribution plots did not provide the same satisfactory insights into the product quality issue (not shown). This may have been due to the fact that the NOC model built on $\mathbf{X}_1$ did not account for all *in-control* sources of variation in $\mathbf{X}_2$, which results in confounding with the *out-of-control* ones in control charts and contribution plots. On the other

[ii]Exploring the distinctive component(s) of $\mathbf{X}_1$ (column of $\mathbf{V}_{1,d}$ according to Section 3.1) might also be of interest in case one wants to investigate aspects like the presence of outlying batches in data supposed to be NOC.

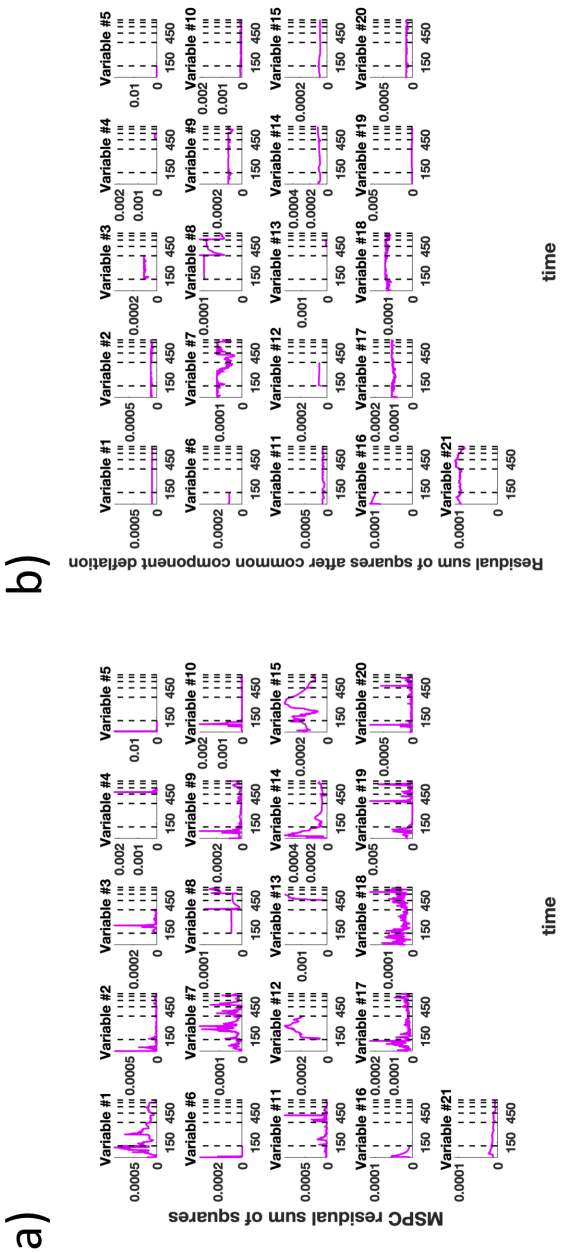

Figure 4: Industrial batch process data - Column-wise residual sum-of-squares (averaged across all the batch runs in $\mathbf{X}_2$) resulting from a) the projection of $\mathbf{X}_2$ onto the *in-control* PCA model calibrated with the data in $\mathbf{X}_1$ and from b) the deflation from $\mathbf{X}_1$ of the common component estimated by the proposed approach. The vertical dashed lines separate the 6 stages of the industrial process. As not all the variables were active in these stages, part of such profiles is missing

hand, fusing the information contained in $\mathbf{X}_1$ and $\mathbf{X}_2$ might have allowed a bigger fraction of the *in-control* variation of $\mathbf{X}_2$ to be modelled and filtered out prior to the exploration of its distinctive components. In order to corroborate this hypothesis, Figures 4a and 4b display the column-wise residual sum-of-squares (averaged across all the batch runs in $\mathbf{X}_2$) resulting from the projection of $\mathbf{X}_2$ onto the *in-control* PCA model calibrated with the data in $\mathbf{X}_1$ and from the deflation from $\mathbf{X}_2$ of the common component estimated by the proposed approach, respectively. As one can clearly see, a larger amount of systematic *in-control* variation of $\mathbf{X}_2$ is removed by the second methodology. For the sake of a fair comparison $\mathbf{X}_1$ and $\mathbf{X}_2$ were auto-scaled and scaled to unit sum-of-squares also when classical MSPC was applied.

3.3. Unit 1/Unit 2 data analysis (second level)

When the analysis was extended to the global ensemble of available data (two sets recorded in the first reactor and two sets recorded in the second reactor, see Figure 1), only a single common *within-distinctive* component was isolated by the CCA-based permutation test (not shown). As expected, owing to the fact that the common component retrieval is attained by the SVD-based modelling technique outlined in Section 3.1, its loadings profiles (first columns of $\mathbf{V}_{1,c}$ and $\mathbf{V}_{2,c}$, respectively, for the second level of the hierarchical structure graphed in Figure 1) are very similar between units, but not identical (see Figures 5a and 5b) [8]. They, however, highlight that variable #1 (starting from the second process stage) features the most consistent contribution to this factor over time. Therefore, this variable could constitute the common problem affecting both reacting units (see also Figure 6 which confirms that variable #1 exhibited a higher average level in the second period batch runs in both reacting units from the second process stage on).

For the sake of completeness, the loading profiles of the first distinctive *between-distinctive* component (statistically significant and corresponding to the first columns of $\mathbf{V}_{1,d}$ and $\mathbf{V}_{2,d}$, respectively, for the second level of the hierarchical structure graphed in Figure 1) are represented in Figure 7. For several patterns of variables (e.g., variables #8, #12, #13, #14, #15 and #16 for the first reactor, and variables #8, #16, #17, #18 and #21 for the second reactor) a consistent non-zero temporal trend is observed. These variables might have been affected by a specific abnormal event occurring in the respective unit. It is also important to notice that variable #1 is characterised by

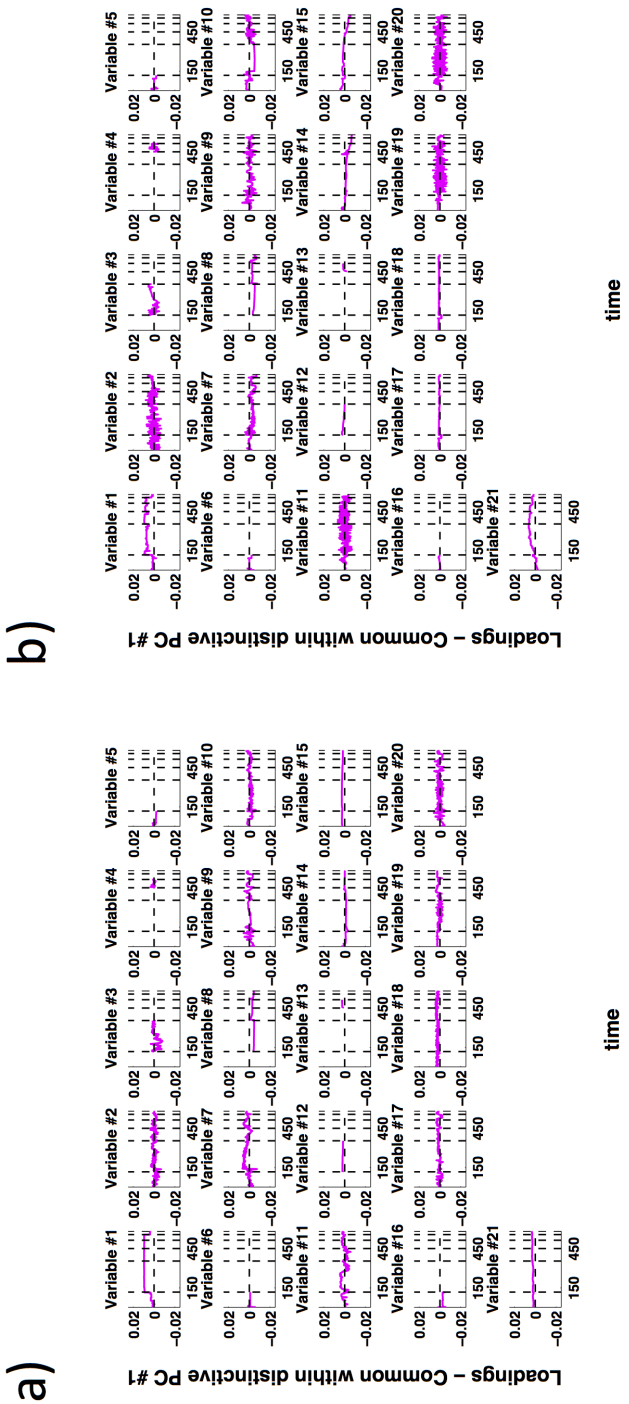

Figure 5: Industrial batch process data - a) Reacting unit 1 *vs* b) reacting unit 2: time profiles of the loadings of the first common *within-distinctive* component for the 21 measured variables (first columns of $\mathbf{V}_{1,c}$ and $\mathbf{V}_{2,c}$, respectively, for the second level of the hierarchical structure graphed in Figure 1). The vertical dashed lines separate the 6 stages of the industrial process. As not all the variables were active in these stages, part of such profiles is missing

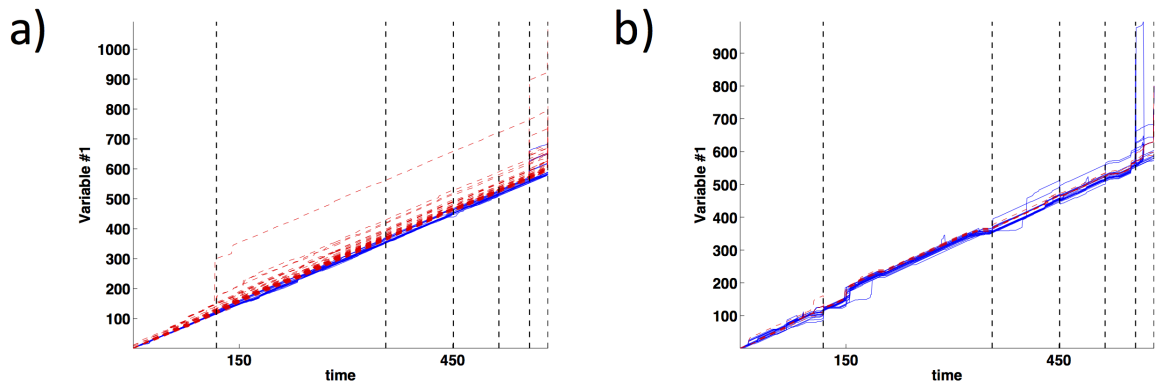

Figure 6: Industrial batch process data - a) Reacting unit 1 *vs* b) reacting unit 2: original time trajectories of variable #1 for the first period (blue solid lines) and the second period (red dashed lines) runs. The vertical dashed lines separate the 6 stages of the industrial process

practically zero and non-consistent loadings from the second process stage on, which is in good agreement with what was stated above for the unique common *within-distinctive* factor.

4. Conclusions

In this article, the exploration of common and distinctive sources of variation in multi-set data was shown to be a promising methodology for industrial batch process troubleshooting and understanding, as an alternative or a complement to classical MSPC. As highlighted in Section 3.2, an accurate MSPC scheme should account for all the *in-control* variation of NOC data in order to be able to correctly assess the quality of future process runs. If particular NOC events which explain very small amounts of such an *in-control* variation are preponderant in new batches, this MSPC scheme could clearly suffer from severe limitations that may be overcome by the approach described here.

Apart from studying data from a single process unit, it also allowed to investigate what distinct time periods in two manufacturing units have in common, which brought to the forefront small sources of variation that would have been otherwise obscured by much larger normal process variability.

A proposal for estimating the number of significant common and distinctive components in multiple blocks of data and modelling them was here presented as well, even though a full evaluation

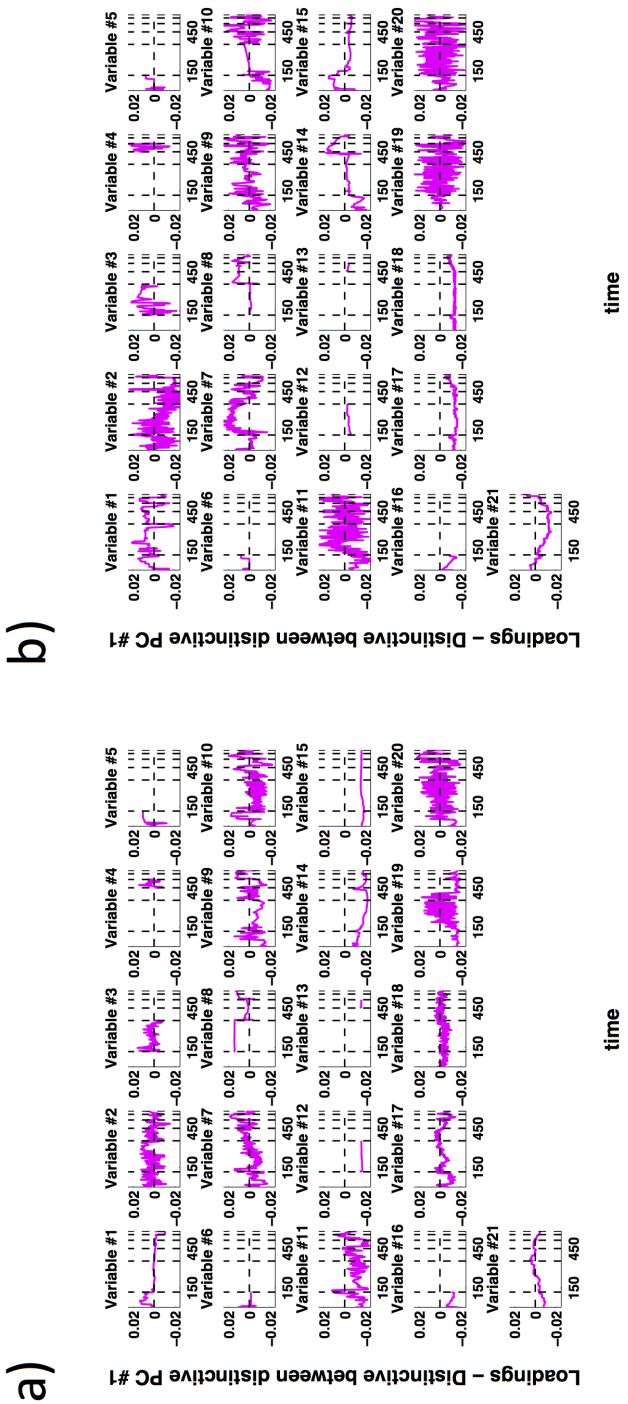

Figure 7: Industrial batch process data - a) Reacting unit 1 *vs* b) reacting unit 2: time profiles of the loadings of the first distinctive *between-distinctive* component for the 21 measured variables (first columns of $\mathbf{V}_{1,\mathrm{d}}$ and $\mathbf{V}_{2,\mathrm{d}}$, respectively, for the second level of the hierarchical structure graphed in Figure 1). The vertical dashed lines separate the 6 stages of the industrial process. As not all the variables were active in these stages, part of such profiles is missing

of its properties will be addressed in future studies. In general, the procedure can be used in a flexible way to focus on the part of the data variation that is most useful for the relevant questions. Its comparison with other methods for common and distinctive component analysis (see Table 1) will be subject of further research.

5. Acknowledgements

This research work was partially supported by the Spanish Ministry of Economy and Competitiveness under the project DPI2017-82896-C2-1-R and Shell Global Solutions International B.V. (Amsterdam, The Netherlands).

6. References

[1] J. Camacho, J. Picó, A. Ferrer, Bilinear modelling of batch processes. Part I: theoretical discussion, J. Chemometr. 22 (2008) 299–308.

[2] J. Camacho, J. Picó, A. Ferrer, Bilinear modelling of batch processes. Part II: a comparison of PLS soft-sensors, J. Chemometr. 22 (2008) 533–547.

[3] J. González-Martínez, J. Camacho, A. Ferrer, Bilinear modelling of batch processes. Part III: parameter stability, J. Chemometr. 28 (2014) 10–27.

[4] T. Kourti, J. MacGregor, Multivariate SPC methods for process and product monitoring, J. Qual. Technol. 28 (1996) 409–428.

[5] P. Nomikos, J. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41–59.

[6] A. Smilde, I. Måge, T. Næs, T. Hankemeier, M. Lips, H. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, J. Chemometr. 31 (2017) e2900.

[7] I. Måge, A. Smilde, F. Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemometr.

[8] K. Van Deun, A. Smilde, L. Thorrez, H. Kiers, I. Van Mechelen, Identifying common and distinctive processes underlying multiset data, Chemometr. Intell. Lab. 129 (2013) 40–51.

[9] H. Kiers, J. ten Berge, Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure, Brit. J. Math. Stat. Psy. 47 (1994) 109–126.

[10] K. Van Deun, A. Smilde, M. van der Werf, H. Kiers, I. Van Mechelen, A structured overview of simultaneous component based data integration, BMC Bioinformatics 10 (2009) 246–260.

[11] M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behav. Res. Methods 45 (2013) 822–833.

[12] K. Van Deun, I. Van Mechelen, L. Thorrez, M. Schouteden, B. De Moor, M. van der Werf, L. De Lathauwer, A. Smilde, H. Kiers, DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes, PLoS One 7 (2012) e37840.

[13] C. Paige, M. Saunders, Towards a generalized singular value decomposition, SIAM J. Numer. Anal. 18 (1981) 398–405.

[14] S. Friedland, A new approach to generalized singular value decomposition, SIAM J. Matrix Anal. A. 27 (2005) 434–444.

[15] M. Schouteden, K. Van Deun, I. Van Mechelen, ECO-POWER: a novel method to reveal common mechanisms underlying linked data, in: Proceedings of the 20th International Conference on Computational Statistics (COMPSTAT 2012), Physica-Verlag, Heidelberg, Germany, 2012, pp. 757–768.

[16] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.

[17] J. Trygg, O2-PLS for qualitative and quantitative analysis in multivariate calibration, J. Chemometr. 16 (2002) 283–293.

[18] E. Lock, K. Hoadley, J. Marron, A. Nobel, Joint and Individual Variation Explained (JIVE) for intergrated analysis of multiple data types, Ann. Appl. Stat. 7 (2013) 523–542.

[19] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1995) 31–58.

[20] J. González-Martínez, O. de Noord, A. Ferrer, Multisynchro: a novel approach for batch synchronization in scenarios of multiple asynchronisms, J. Chemometr. 28 (2014) 462–475.

[21] J. González-Martínez, R. Vitale, O. de Noord, A. Ferrer, Effect of synchronization on bilinear batch process modeling, Ind. Eng. Chem. Res. 53 (2014) 4339–4351.

[22] I. Måge, B. Mevik, T. Næs, Regression models with process variables and parallel blocks of raw material measurements, J. Chemometr. 22 (2008) 443–456.

[23] R. Vitale, J. Westerhuis, T. Næs, A. Smilde, O. de Noord, A. Ferrer, Selecting the number of factors in principal component analysis by permutation testing - numerical and practical aspects, J. Chemometr. 31 (2017) e2937.

[24] M. Timmerman, H. Kiers, Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences, Psychometrika 68 (2003) 105–121.

[25] H. Kiers, A. Smilde, A comparison of various methods for Multivariate Regression with highly collinear variables, Stat. Method. Appl. 16 (2007) 193–228.

[26] R. Van den Berg, C. Rubingh, J. Westerhuis, M. van der Werf, A. Smilde, Metabolomics data exploration guided by prior knowledge, Anal. Chim. Acta 651 (2009) 173–181.

[27] T. Dahl, T. Næs, A bridge between Tucker-1 and Carroll's generalized canonical analysis, Comput. Stat. Data An. 50 (2006) 3086–3098.

[28] A. Tenenhaus, M. Tenenhaus, Regularized Generalized Canonical Correlation Analysis, Psychometrika 76 (2011) 257–284.

Appendix A. Canonical Correlation Analysis (CCA)

Let $\mathbf{X}_k$ be the k-th of multiple data matrices with dimensions $N \times J_k$ in the object-wise linked data case. Assume from now on that each $\mathbf{X}_k$ is initially auto-scaled[iii] and afterwards scaled to equal sum-of-squares[iv].

[iii] Auto-scaling corresponds to centering and scaling to unit variance the concerned J_k or J variables.

[iv] This will allow all the measured variables to have equal weight and prevent potential bias due to differences in e.g., the size of the various $\mathbf{X}_k$ [22, 24].

Canonical Correlation Analysis (CCA) [16] is a technique suitable for handling pairs of object-wise linked datasets. Here, the $\mathbf{X}_k$ blocks ($k = 1, 2$) are modelled as:

$$\mathbf{X}_k = \mathbf{X}_k\mathbf{W}_k\mathbf{P}_k^{\mathrm{T}} + \mathbf{E}_k = \mathbf{T}_k\mathbf{P}_k^{\mathrm{T}} + \mathbf{E}_k \tag{A.1}$$

where $\mathbf{W}_k$ ($J_k \times A$) is a matrix containing the so-called canonical weights, $\mathbf{T}_k$ ($N \times A$) represents the canonical variate array, while the loadings $\mathbf{P}_k$ ($J_k \times A$) are obtained by regressing $\mathbf{X}_k$ on $\mathbf{T}_k = \mathbf{X}_k\mathbf{W}_k$. CCA solves the following objective function:

$$\max_{\mathbf{W}_1,\mathbf{W}_2} \mathrm{tr}(\mathbf{W}_1^{\mathrm{T}}\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_2\mathbf{W}_2) \quad \text{s.t. } N^{-1}\mathbf{T}_1^{\mathrm{T}}\mathbf{T}_1 = \mathbf{I} = N^{-1}\mathbf{T}_2^{\mathrm{T}}\mathbf{T}_2 \tag{A.2}$$

Thus, $\mathbf{W}_1$ and $\mathbf{W}_2$ result from the maximisation of the sum of the correlations between the A couples of canonical variates. Since the variance of the different $\mathbf{X}_k$ explained by such canonical variates is not taken into account in Equation A.2, they might be poor descriptors of the original data [25]. In order to overcome this limitation, which can generate certain instability in the final outcomes[v], one may apply PCA block-wise prior to CCA or use regularisation [26–28]. The extension of the CCA algorithm for coping with more than two datasets is known as Generalised Canonical Correlation Analysis (GCCA) [28].

Appendix B. Effective rank determination algorithm

Let $\mathbf{X}$ be a centred data matrix of N rows and J columns with rank $Q = \min\{N - 1, J\}$. The novel computational procedure proposed in [23] comprises the following 10 steps grouped in three consecutive phases:

Phase I - Singular Value Decomposition of $\mathbf{X}$:

1. Perform Singular Value Decomposition (SVD) on $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \tag{B.1}$$

where $\mathbf{U}$ ($N \times N$) and $\mathbf{V}$ ($J \times J$) contain the left and right singular vectors of $\mathbf{X}$, respectively, and $\mathbf{S}$ ($N \times J$) is a rectangular diagonal array whose non-zero diagonal elements are its singular values ($\sqrt{\lambda_1}$, $\sqrt{\lambda_2}, \ldots,$ $\sqrt{\lambda_Q}$);

[v] e.g., when $N < J_k$ for some k Equation A.2 leads to an undetermined system of equations.

2. Compute for each a-th calculated component the ratio:

$$F_a = \frac{\lambda_a}{\sum_{q=a}^{Q} \lambda_q} \tag{B.2}$$

where λ_a corresponds to the a-th eigenvalue obtained after the decomposition of $\mathbf{X}$. F_a is used for testing the statistical significance of the single factors. It equals the ratio between the amount of variation explained by the a-th component and the total amount of variation captured by the last $Q - (a - 1)$ components.

Phase II - Test for the first component:

3. For $a = 1$, randomly and independently permute the order of the entries within every column of $\mathbf{X}$ constructing a new matrix $\mathbf{X}_{perm}$, featuring uncorrelated variables;
4. Apply SVD to $\mathbf{X}_{perm}$ and calculate the ratio:

$$F_{1,perm} = \frac{\lambda_{1,perm}}{\sum_{q=1}^{Q} \lambda_{q,perm}} \tag{B.3}$$

where $\lambda_{1,perm}$ denotes the first eigenvalue obtained after the decomposition of $\mathbf{X}_{perm}$. Note that the sum of squares of $\mathbf{X}$ and $\mathbf{X}_{perm}$ is exactly the same, despite the permutations;

5. Iterate step 3 and 4 to generate a *null*-distribution for $F_{1,perm}$[vi]. If F_1 is found to be higher than its $(1 - \alpha) \times 100^{\text{th}}$ percentile (α equals the nominal Overall Type I - OTI - risk value imposed to the test, i.e., its false positive rate), the first component is considered statistically significant.

Phase III - Test for the a-th component ($a > 1$):

6. For $a > 1$, calculate the residual matrix:

$$\mathbf{E}_a = \mathbf{X} - \sum_{q=1}^{a-1} \mathbf{u}_q \sqrt{\lambda_q} \mathbf{v}_q^{\mathrm{T}} = \mathbf{X} - \sum_{q=1}^{a-1} \mathbf{t}_q \mathbf{p}_q^{\mathrm{T}} \tag{B.4}$$

[vi]The total number of iterations is a user-defined parameter and should be selected so as to obtain a precise estimation of such a *null*-distribution.

where $\mathbf{u}_q$, $\mathbf{v}_q$, $\mathbf{t}_q$ and $\mathbf{p}_q$ are the q-th column vectors of $\mathbf{U}$, $\mathbf{V}$, $\mathbf{T}$ and $\mathbf{P}$ (see Equation B.1), respectively[vii]. Note that after each deflation round $\mathbf{E}_a$ has rank $Q-(a-1)$;

7. Randomly and independently permute the order of the entries within each column of $\mathbf{E}_a$ constructing a new matrix $\mathbf{E}_{a,perm}$. Unlike $\mathbf{E}_a$, $\mathbf{E}_{a,perm}$ has rank Q (apart from chance deviations), but their total sums of squares are the same;
8. Calculate the projection of $\mathbf{E}_{a,perm}$ on a subspace of dimensionality $Q-(a-1)$, $\mathbf{E}_{a,perm,\text{proj}}$, as:

$$\mathbf{E}_{a,perm,\text{proj}} = (\mathbf{I}_N - \sum_{q=1}^{a-1} \mathbf{u}_{q,\mathbf{E}_{a,perm}} \mathbf{u}_{q,\mathbf{E}_{a,perm}}^{\mathrm{T}}) \mathbf{E}_{a,perm} \tag{B.5}$$

where $\mathbf{I}_N$ is an identity matrix of dimensions $N \times N$, $\mathbf{E}_{a,perm} = \mathbf{U}_{\mathbf{E}_{a,perm}} \mathbf{S}_{\mathbf{E}_{a,perm}} \mathbf{V}_{\mathbf{E}_{a,perm}}^{\mathrm{T}}$, and $\mathbf{u}_{q,\mathbf{E}_{a,perm}}$ is the q-th column vector of $\mathbf{U}_{\mathbf{E}_{a,perm}}$;

9. Perform SVD on $\mathbf{E}_{a,perm,\text{proj}}$ and retain the ratio:

$$F_{a,perm,\text{proj}} = \frac{\lambda_{1,perm,\text{proj}}}{\sum_{q=1}^{Q-(a-1)} \lambda_{q,perm,\text{proj}}} \tag{B.6}$$

where $\lambda_{1,perm,\text{proj}}$ is the first eigenvalue obtained after the decomposition of $\mathbf{E}_{a,perm,\text{proj}}$;

10. Iterate step 7, 8 and 9 to generate a *null*-distribution for $F_{a,perm,\text{proj}}$[iv]. If F_a is found to be higher than its $(1-\alpha) \times 100^{\text{th}}$ percentile, the a-th component is considered statistically significant.

Computations are stopped as soon as the first non-significant component is detected.

[vii] According to this notation a hypothetical $\mathbf{E}_1$ would correspond to $\mathbf{X}$.