Chemometrics and Intelligent Laboratory Systems 208 (2021) 104212

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

All sparse PCA models are wrong, but some are useful. Part II: Limitations and problems of deflation

J. Camacho [a,*], A.K. Smilde [b], E. Saccenti [c], J.A. Westerhuis [b], Rasmus Bro [d]

[a] Department of Signal Theory, Telematics and Communications, School of Computer Science and Telecommunications - CITIC, University of Granada, Granada, Spain
[b] Biosystems Data Analysis, University of Amsterdam, Amsterdam, the Netherlands
[c] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, the Netherlands
[d] Chemometrics and Analytical Technology, University of Copenhagen, Denmark

ARTICLE INFO

Keywords:
Artifacts
Data interpretation
Exploratory data analysis
Model interpretation
Sparse principal component analysis
Sparsity

ABSTRACT

Sparse Principal Component Analysis (sPCA) is a popular matrix factorization approach based on Principal Component Analysis (PCA). It combines variance maximization and sparsity with the ultimate goal of improving data interpretation. A main application of sPCA is to handle high-dimensional data, for example biological *omics* data. In Part I of this series, we illustrated limitations of several state-of-the-art sPCA algorithms when modeling noise-free data, simulated following an exact sPCA model. In this Part II we provide a thorough analysis of the limitations of sPCA methods that use deflation for calculating subsequent, higher order, components. We show, both theoretically and numerically, that deflation can lead to problems in the model interpretation, even for noise free data. In addition, we contribute diagnostics to identify modeling problems in real-data analysis.

1. Introduction

The characteristics and the properties of algorithms to perform sparse principal component analysis (sPCA) have been discussed in the statistical and data analysis literature [1–4]. However, the derivation of multi-component sPCA models, with two or more components, has received limited attention.

This is the second paper of a series devoted to the critical assessment of several widely used multi-component sPCA algorithms. In the first paper [5], we showed that, under specific circumstances, many popular sPCA variants [2,6] are unable to model noise-free sparse data accurately. In this Part II, we provide a theoretical explanation for this behavior, supported by Monte Carlo simulations, and propose diagnostic statistics to identify modeling problems in real-data analysis. We focus on approaches based on deflation, notably the Penalized Matrix Decomposition (PMD) by Witten et al. [2] and the Group-wise PCA (GPCA) by Camacho et al. [7]. We show that complex interaction among components can lead to problems for model interpretation in sPCA.

The rest of the paper is organized as follows. Section 2 introduces the use of deflation-based models in sPCA. Section 3 discusses sPCA algorithms that are based on deflation. Section 4 describes the introduction of artifacts in multi-component models based on deflation. Section 5 presents the materials and methods for the experimental part of this work. Section 6 quantifies the artifacts using simulated and real data and Section 7 draws conclusions.

2. Sequential models based on deflation

A widely used approach in multi-component algorithms is to fit a sequential series of single components. Often, this is done by deflating each component from the data, in order to compute residuals to fit the next component. Let us define $\mathbf{X}_A (N \times M)$ as the residual matrix after $A-1$ components have been extracted from $\mathbf{X}_1$, which indicates the original matrix of data. The basic equation for deflating $\mathbf{X}_A$ is:

$$\mathbf{X}_{A+1} = \mathbf{X}_A - \mathbf{t}_A \mathbf{p}_A^{\mathrm{T}}, \tag{1}$$

where $\mathbf{t}_A$ and $\mathbf{p}_A$ are vectors of conformable size, representing the scores and loadings for component A.

There are a number of matters regarding Eq. (1) that require investigations and that depend on how $\mathbf{t}_A$ and $\mathbf{p}_A$ were estimated, mainly:

a) In which sub-spaces are the different vectors located?
b) Can explained variances be calculated in a meaningful way?
c) What are the orthogonality properties of the deflation?
d) What are the rank reducing properties of the deflation?

* Corresponding author.
E-mail address: josecamacho@ugr.es (J. Camacho).

https://doi.org/10.1016/j.chemolab.2020.104212
Received 20 July 2020; Received in revised form 13 October 2020; Accepted 26 November 2020
Available online 26 December 2020
0169-7439/

In this paper, we put special emphasis on the study of the sub-spaces of the estimated parameters in connection with the deflation and its orthogonality properties. The computation of explained variance in sPCA was one of the topics in the first part of the series [5], and the rank reducing properties are not treated here.

Deflation in sPCA can be complex, as shown by Mackey [1]. Deflation approaches that work properly in PCA do not necessarily do so in sPCA. Two popular forms of deflation in sPCA are projection deflation and the deflation proposed by Mackey. Projection deflation and Mackey's deflation share similar properties. In the rest of the present paper, we will generally refer to both projection deflation and Mackey's deflation as "loading deflation".

There exist alternative, seldom used, strategies for sPCA deflation or orthogonalization that present different properties from loading deflation: these include the orthogonalization approach in PMD [2], the use of scores for deflation, or the projection of the sparse loadings to the row-space for subsequent deflation. The study of those methods is outside of the scope of this paper.

2.1. Projection deflation

Considering an arbitrary vector $\mathbf{p}_A \in \Re^M$, the data matrix $\mathbf{X}_A$ can be deflated for $\mathbf{p}_A$ with Eq. (1) using a projection step, so that

$$\mathbf{t}_A = \mathbf{X}_A \mathbf{p}_A (\mathbf{p}_A^{\mathrm{T}} \mathbf{p}_A)^{-1}. \tag{2}$$

The matrix $\mathbf{X}_{A+1}$ can be explicitly written:

$$\mathbf{X}_{A+1} = \mathbf{X}_A (\mathbf{I} - \mathbf{p}_A (\mathbf{p}_A^{\mathrm{T}} \mathbf{p}_A)^{-1} \mathbf{p}_A^{\mathrm{T}}). \tag{3}$$

The matrix $\mathbf{I} - \mathbf{p}_A (\mathbf{p}_A^{\mathrm{T}} \mathbf{p}_A)^{-1} \mathbf{p}_A^{\mathrm{T}}$, where $\mathbf{I}$ is the identity matrix of appropriate size, projects orthogonally the rows of $\mathbf{X}_A$ onto the space orthogonal to $\mathbf{p}_A$. A property of this solution is that $\mathbf{X}_{A+1}\mathbf{p}_A = 0$, which is due to the orthogonal regression step. The projection deflation is a form o[1]f orthogonalization.[1]

2.2. Mackey's deflation

Mackey's deflation follows a sequential Gram-Schmidt decomposition:

$$\mathbf{q}_A = \mathbf{B}_A \mathbf{p}_A, \tag{4}$$

$$\mathbf{q}_A = \mathbf{q}_A \Big/ \sqrt{\mathbf{q}_A^{\mathrm{T}} \mathbf{q}_A}, \tag{5}$$

$$\mathbf{X}_{A+1} = \mathbf{X}_A (\mathbf{I} - \mathbf{q}_A \mathbf{q}_A^{\mathrm{T}}), \tag{6}$$

$$\mathbf{B}_{A+1} = \mathbf{B}_A (\mathbf{I} - \mathbf{q}_A \mathbf{q}_A^{\mathrm{T}}), \tag{7}$$

where $\mathbf{B}_1$ is initialized to the identity matrix of appropriate dimension. For the first component, $\mathbf{q}_1$ equals the normalized sparse loading $\mathbf{p}_1$. For subsequent components, $\mathbf{q}_A$ is computed to be in the space orthogonal to all previous sparse loadings $\{1, \ldots, A-1\}$. Mackey's deflation can be re-expressed as:

$$\mathbf{t}_A = \mathbf{X}_A \mathbf{q}_A, \tag{8}$$

and

$$\mathbf{X}_{A+1} = \mathbf{X}_A - \mathbf{t}_A \mathbf{q}_A^{\mathrm{T}}. \tag{9}$$

[1] Orthogonalization can also be done in the other mode: $\mathbf{p}_A^{\mathrm{T}} = (\mathbf{t}_A^{\mathrm{T}} \mathbf{t}_A)^{-1} \mathbf{t}_A^{\mathrm{T}} \mathbf{X}_A$. Then it can be shown that $\mathbf{t}_A$ is orthogonal to $\mathbf{X}_{A+1}$ ($\mathbf{X}_{A+1}^{\mathrm{T}} \mathbf{t}_A = 0$).

For the first component, Eqs. (8) and (9) in Mackey's deflation equal Eqs. (2) and (1) in projection deflation, provided $\mathbf{p}_1$ is of unit length in the latter.

3. Sparse PCA algorithms using loading deflation

We consider two sPCA algorithms based on loading deflation: the Penalized Matrix Decomposition (PMD) by Witten et al. [2] and the Group-wise PCA (GPCA) by Camacho et al. [7].

3.1. Penalized Matrix Decomposition

The PMD extends a rank one Singular Value Decomposition (SVD) by considering sparsity constraints in both the right and left vectors in the matrix factorization. Restricting this framework to constrain only the loadings, the sparse PCA problem is formulated as:

$$\{\hat{\mathbf{p}}_A^P, \hat{\mathbf{u}}_A^P\} = argmax_{\mathbf{p}_A, \mathbf{u}_A} \ \mathbf{u}_A^{\mathrm{T}} \mathbf{X}_A \mathbf{p}_A \quad s.t. \quad \|\mathbf{p}_A\|_1 \le c_2, \quad \|\mathbf{p}_A\|_2^2 \le 1, \quad \|\mathbf{u}_A\|_2^2 \le 1, \tag{10}$$

where the superscript P refers to *PMD* and $\hat{\mathbf{p}}_A^P$ is obtained using a soft-thresholding operator and subsequently $\hat{\mathbf{u}}_A^P$ is obtained by a least squares operation. The corresponding pseudo-singular value is obtained as:

$$\hat{d}_A^P = (\hat{\mathbf{u}}_A^P)^{\mathrm{T}} \mathbf{X}_A \hat{\mathbf{p}}_A^P. \tag{11}$$

After the A-th component is obtained, the deflation is performed as:

$$\mathbf{X}_{A+1} = \mathbf{X}_A - \hat{d}_A^P \hat{\mathbf{u}}_A^P (\hat{\mathbf{p}}_A^P)^{\mathrm{T}}. \tag{12}$$

Since $\hat{\mathbf{u}}_A^P$ and $\hat{d}_A^P$ are obtained by least squares (see Appendix A for a proof), Eq. (12) equals the projection deflation discussed before.

3.2. Group-wise PCA

While most sPCA algorithms modify the classical PCA by including sparsity-inducing constraints or penalties with the L_0 or L_1 norms [8], Group-wise PCA uses a completely different approach. It starts with the identification of a set of G (possibly overlapping) groups of correlated variables. Then, the GPCA algorithm computes G candidate loading vectors, where the g-th vector can only have non zero elements associated to the variables in the g-th group. From these G candidates, only the one that has the largest variance of the component, that is arg max$p$ ($bp^{\mathrm{T}} \mathbf{X}_A^{\mathrm{T}} \mathbf{X}_A \mathbf{p}$) for $\mathbf{p}$ of unit length, is retained and it is used to deflate data matrix $\mathbf{X}_A$ using the deflation by Mackey.

4. Deflation in sparse PCA can produce artifacts

In this section, we discuss situations in which the deflation mechanism can introduce artifacts or unexpected variance in sPCA models, which may hamper their interpretation. Problems may arise depending on the relationship among the sparse loadings, and with the data row-space, so that we have basically two considerations to make:

- Are all loadings orthogonal?
- Are all loadings in the data row-space?

If the answer to the first question is no, the sparse components from the second onwards may be difficult to interpret or even misleading, since complex interactions among components are generated by the deflation mechanism. This is especially the case when the answer to the second question is also no, and sparse loadings are outside the data row-space.

In the following, we will start by describing the concepts of data column/row-spaces and their relevance in the sPCA context. Then we

will proceed with the description of the situation when loadings are out of the data row-space and non-orthogonal, which represents the most challenging but common case. Finally, we will address the case when loadings are inside the data row-space but non-orthogonal.

4.1. Sparse vectors and the data column and row-spaces

The column-space $R(\mathbf{X}_1)$ of an $N \times M$ matrix $\mathbf{X}_1$ is the span of the columns of $\mathbf{X}_1$, hence all vectors $\mathbf{b} \in \mathbb{R}^N$ which can be written as $\mathbf{b} = \mathbf{X}_1\mathbf{a}$ for arbitrary $\mathbf{a} \in \mathbb{R}^M$. Similarly, the row-space of $\mathbf{X}_1$ are all the vectors in the form $\mathbf{d} = \mathbf{X}_1^{\mathrm{T}}\mathbf{c}$, for $\mathbf{d} \in \mathbb{R}^M$ and $\mathbf{c} \in \mathbb{R}^N$, and it is denoted by $R(\mathbf{X}_1^{\mathrm{T}})$. It is a well-known property of a PCA of $\mathbf{X}_1$ that the resulting scores and loadings are in the column, respectively row-space, of $\mathbf{X}_1$ [9].

Let us focus on the row-space $R(\mathbf{X}_1^{\mathrm{T}})$ since in most cases constraints to obtained sparse PCA solutions are imposed on the loadings. Consider two cases:

1. Low-dimensional case, when $M \leq N$
2. High-dimensional case, when $M > N$

Since $\mathbf{X}_1$ contains measured data, due to the presence of measurement noise, we assume $\mathbf{X}_1$ to be of full rank, meaning the rank equals the size of the smallest dimension.

In the low-dimensional case, the rank of $\mathbf{X}_1$ is M and thus $R(\mathbf{X}_1^{\mathrm{T}}) = \mathbb{R}^M$. Hence, all vectors of size M in $\mathbb{R}^M$ are in the row-space of $\mathbf{X}_1$, and this includes also sparse loadings. In the high-dimensional case, the rank of $\mathbf{X}_1$ is N and $R(\mathbf{X}_1^{\mathrm{T}}) \subset \mathbb{R}^M$ and, as a consequence, some vectors of size M are in the row-space of $\mathbf{X}_1$ and some are not. In this situation, sparse loadings can be outside the row-space of $\mathbf{X}_1$ as a result of constraints/penalties implemented to arrive at sparse solutions. See a proof in Appendix B.

From Eqs. (2) and (8), it also follows that $\mathbf{t}$ in sPCA is always in the column-space of $\mathbf{X}_1$.

4.2. Non-orthogonal loadings out of the data row-space

As already discussed, the sparse loadings in sPCA can be outside the row-space of the data, $\mathbf{X}_1$, from which they are fitted. As a matter of fact, this is expected to happen in any data set with measurement noise. It should be noted that going outside the original data space is an expected consequence of using any form of regularization or constraint in the model, since this modifies the data fitting towards some pre-imposed *a priori* structure (such as sparsity). Provided that these *a priori* assumptions are realistic, and data support them to a reasonable level, going outside the data space is not a problem.

A sparse component will get outside the row-space $R(\mathbf{X}_1^{\mathrm{T}})$ when $M > N$ and $\mathbf{X}_1$ itself is not sparse, which is the case when data contains noise (likely the most general case) or because we employ sPCA as a means to simplify the interpretation in non-sparse data. Even when $\mathbf{X}_1$ is sparse, the component can get outside the data row-space (we will see this situation in the simulation examples).

When a loading $\mathbf{p}_1$ is outside $R(\mathbf{X}_1^{\mathrm{T}})$ and $\mathbf{X}_2$ is obtained from loading deflation (either projection or Mackey's deflation), it holds that $R(\mathbf{X}_2^{\mathrm{T}}) \nsubseteq R(\mathbf{X}_1^{\mathrm{T}})$. Hence, we introduce new directions (artifacts) in $\mathbf{X}_2$ which were not present in the original data $\mathbf{X}_1$. This may be undesirable because these artifacts might end up in components obtained from $\mathbf{X}_2$, and thus lead to interpretations of the data that are incorrect.

To find the discrepancy between residuals after the deflation of $A-1$ components, $\mathbf{X}_A$, and the original data $\mathbf{X}_1$, the residuals $\mathbf{O}_A$ of projecting $R(\mathbf{X}_A^{\mathrm{T}})$ onto $R(\mathbf{X}_1^{\mathrm{T}})$ can be considered:

$$\mathbf{O}_A = (\mathbf{I} - \mathbf{X}_1^{\mathrm{T}}(\mathbf{X}_1^{\mathrm{T}})^{+})\mathbf{X}_A^{\mathrm{T}}, \tag{13}$$

where $(\mathbf{I} - \mathbf{X}_1^{\mathrm{T}}(\mathbf{X}_1^{\mathrm{T}})^{+})$ describes the projection to the space orthogonal to the row-space of $\mathbf{X}_1$. We refer to $\mathbf{O}_A$ as the spurious residuals. If they are not null, this means artifacts were introduced in the residuals $\mathbf{X}_A$ after deflation due to the departure of the row-space.

However, the fact that spurious residuals exist does not mean that subsequent components are affected by artifacts. To measure the amount of artifacts, that is, the amount of variance in a component that can not be attributed to the original data and must hence be considered spurious, we can use the following expression:

$$\hat{\mathbf{O}}_{A+1} = (\mathbf{I} - \mathbf{X}_1^{\mathrm{T}}(\mathbf{X}_1^{\mathrm{T}})^{+})(\mathbf{X}_A^{\mathrm{T}}(\mathbf{X}_A^{\mathrm{T}})^{+})\mathbf{p}_{A+1}\mathbf{t}_{A+1}^{\mathrm{T}}. \tag{14}$$

Eq. (14) answers to the question of how much variance of component $\mathbf{p}_{A+1}$, represented by $\mathbf{p}_{A+1}\mathbf{t}_{A+1}^{\mathrm{T}}$, lies in the row-space of $\mathbf{X}_A$, $(\mathbf{X}_A^{T}(\mathbf{X}_A^{T})^{+})$, and not in the row-space of $\mathbf{X}_1$, $(\mathbf{I} - \mathbf{X}_1^{T}(\mathbf{X}_1^{T})^{+})$. This can be interpreted as how much of $\mathbf{p}_{A+1}$ is actually built from the spurious residuals in Eq. (13), and therefore correspond to artifacts. A detailed description of the rationale behind Eq. (14) can be found in Appendix B.3.

The percentage of artifacts of the component $A+1$ can be computed as:

$$VarA_{A+1} = 100 \times \frac{tr(\hat{\mathbf{O}}_{A+1}^{\mathrm{T}}\hat{\mathbf{O}}_{A+1})}{tr(\mathbf{p}_{A+1}\mathbf{t}_{A+1}^{\mathrm{T}}\mathbf{t}_{A+1}\mathbf{p}_{A+1}^{\mathrm{T}})}. \tag{15}$$

This is a useful diagnostic statistic, as we will illustrate in the experiments, to identify the amount of artifacts contaminating components. If the percentage of artifacts is large, care should be taken to understand how these would affect the interpretation component $A+1$.

4.3. Non-orthogonal loadings in the data row-space

Even for the unrealistic situation of data generated with a perfect (noise-less) multi-component sPCA model, if the sparse loadings overlap, sPCA may still be difficult to interpret. We say two sparse loadings overlap when they have non-zero loadings in at least one common variable. A multi-component sPCA model with overlapping components inherits from PCA the rotational ambiguity: we can rotate several loadings maintaining the same performance in the loss function. A trivial case to see this is when two sparse loadings have their non-zero elements in the same, exact variables. Then, they can be rotated with orthogonal rotation methods within the space of those variables, and the model loss will remain the same. The study of rotational ambiguity in sPCA models is outside of the scope of this paper, but it deserves future attention.

On the other hand, again in the optimistic situation of data generated with a perfect (noise-less) multi-component sPCA model, even if a sparse loading accurately matches a true component, loading deflation can produce a transfer of variance among components that may affect interpretation. This was observed for the first time in the case of multiblock PLS [10]. We will make use of a toy example to illustrate the problem. Let us take a data set $\mathbf{X}_1$ following a perfect sPCA model:

$$\mathbf{X}_1 = \mathbf{T}\mathbf{P}^{\mathrm{T}}, \tag{16}$$

with:

$$\mathbf{T} = 1/c \cdot [\mathbf{x}_a,\ \mathbf{x}_b], \tag{17}$$

and

$$\mathbf{P} = \begin{bmatrix} c & c & 0 \\ 0 & c & c \end{bmatrix}^{\mathrm{T}}, \tag{18}$$

where for convenience $\mathbf{x}_a$ and $\mathbf{x}_b$ are columns vectors *i.i.d.* following a standard normal distribution and $c = \sqrt{2}/2$. This results in:

$$\mathbf{X}_1 = [\mathbf{x}_a,\ \mathbf{x}_a + \mathbf{x}_b,\ \mathbf{x}_b]. \tag{19}$$

To give a little of context to this example, imagine that data comes from a treatment-control experiment, that $\mathbf{x}_a$ represents uninteresting variance while $\mathbf{x}_b$ contains the level of disease, and that we are looking

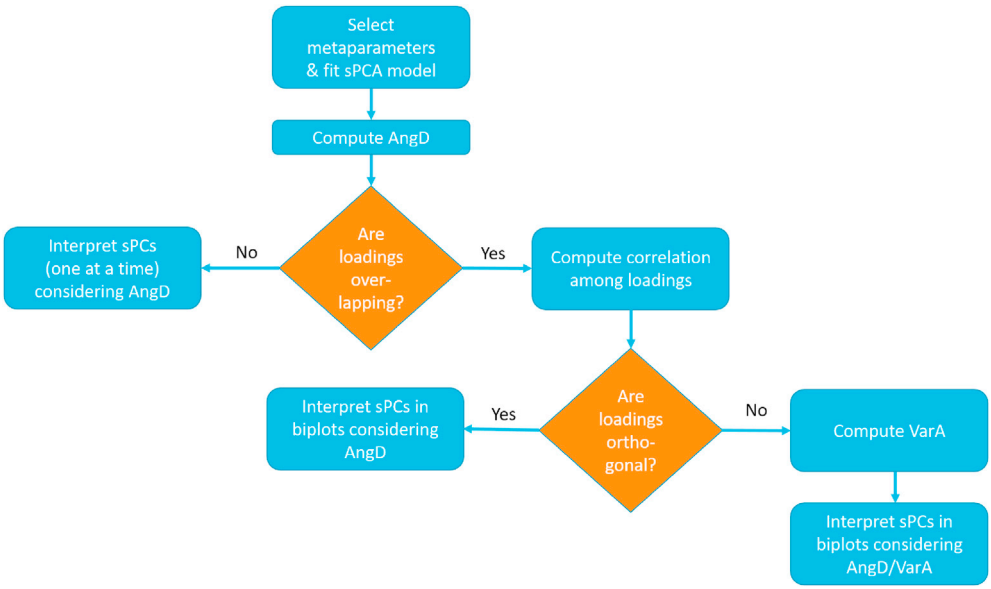

Fig. 1. Interpretation of sPCA models based on loading deflation.

for biomarkers for the effectiveness of that treatment. Therefore, the second and third variables in $\mathbf{X}_1$ can be potential biomarkers since they contain $\mathbf{x}_b$.

It can be seen that the first loading $\mathbf{p}_1 = [c, c, 0]^{\mathrm{T}}$ belongs to $R(\mathbf{X}_1^{\mathrm{T}})$, and therefore we are not in the situation described in the previous subsection, where loadings were outside the data row-space. Let us perform the deflation over $\mathbf{X}_1$ using the real loading. First, we estimate the scores with Eq. (2) and residuals with Eq. (3):

$$\mathbf{t}_1 = 2c \cdot \mathbf{x}_a + c \cdot \mathbf{x}_b, \tag{20}$$

and

$$\mathbf{X}_2 = \mathbf{X}_1 - [c \cdot \mathbf{t}_1,\ c \cdot \mathbf{t}_1,\ 0]. \tag{21}$$

Operating:

$$\mathbf{X}_2 = [1/2 \cdot \mathbf{x}_b,\quad 1/2 \cdot \mathbf{x}_b,\quad \mathbf{x}_b]. \tag{22}$$

Notice this derivation of the residuals of the first component is equivalent with Mackey's deflation.

Eq. (22) shows that $\mathbf{x}_b$ has been transferred to the first column of the deflated data, when it was not originally there. This affects the computation of subsequent sparse components, and their interpretation may be misleading, since it can lead us to think that the first variable is related to the disease of interest.

Generally speaking, this transfer effect introduces variance of some components (or of some groups of variables) in others in the deflation process, like it happens among blocks in multi-block models [10]. Like the artifacts in the previous section, this transfer only happens for overlapping sparse loadings, and in particular when loadings are non-orthogonal.

In principle, given that in this case the artifacts are not the result of departing from the data row-space, we cannot detect this problem using the diagnostic statistic in Eq. (15). While this situation is highly unlikely in real practice, the take-away message is that care should be taken for over-interpreting sPCA models where loadings in different components overlap.

4.4. Identification of problems in practical situations

To quantify the problems deriving from the limitations in loading deflation sPCA, we define a set of diagnostic statistics. These statistics are intended to provide a measure of the quality of the model in practical applications with respect to the amount of distortion/artifacts introduced, and are defined as follows:

- AngD_A: The angle (in degrees) of the A-th sparse loading to the row-space of $\mathbf{X}_1$. This parameter is between 0 and 90°.
- VarA_A: The percentage of captured variance of the A-th component that lies outside the data row-space $R(\mathbf{X}_1^{\mathrm{T}})$ according to Eq. (15). This parameter is between 0 and 100.
- RSS: The residual sum-of-squares normalized by the total sum-of-squares of the data according to Eq. (23). This parameter is often used as a model goodness criteria in sPCA [6,11].

$$RSS = \frac{\mathrm{tr}(\mathbf{E}^{\mathrm{T}}\mathbf{E})}{\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X})} \tag{23}$$

While the RSS is a commonly used statistic in sPCA, as far as we know, the other two statistics have not been used before in this context.

In Fig. 1, we propose a number of steps intended to safely interpret sPCA models obtained with loading deflation. The first step is to select the number of components and non-zero elements following established strategies, see for instance Ref. [2,12]. Subsequently, the AngD_A statistic is computed per component. Components which loadings do not overlap with any other can be interpreted individually, using e.g. one-component bar charts of scores and loadings. Care should be taken with components of very high AngD_A,[2] since this may spot situations in which the sparsity induced by the model is not supported by the data. For overlapping loadings, the correlation should be computed. If the loadings of two components are orthogonal, traditional score and loading plots can be safely used for interpretation, again taking the AngD_A values into consideration. If loadings are not orthogonal, we are in the situation of higher risk of artifacts. Therefore, we suggest to compute the Var_A statistic, and only trust the interpretation of components where artifacts remain reasonably low.[3]

[2] In our experiments, values of AngD_A below 20ordm; seem reasonable, but more research is needed to determine reasonable thresholds for different data structures.

[3] Again, in our experiments, values of Var_A below 20% seem reasonable, but more research is needed also in this point.

5. Material and methods

The numerical results in this paper are based on a number of simulations and several variants of the sPCA algorithms outlined in Section 3. Simulations and algorithms are inherited from the first paper of the series [5] to which we refer the reader for more details.

5.1. Algorithms for sparse PCA[2]

The original definitions of PMD [2] and GPCA [7] differ in how they fit components, as discussed in Section 3, and in how they deflate, PMD using projection deflation and GPCA Mackey's deflation. These steps can be interchanged: we can use the PMD optimization function (10) coupled with Mackey's generalized deflation and GPCA with the projection deflation. Thus, we consider the following variants of sPCA based on deflation:

1. The PMD algorithm [2] with projection deflation (PMD-PD) and a modified version using Mackey's generalized deflation (PMD-M).
2. The GPCA algorithm [7] using projection deflation (GPCA-PD) and Mackey 's generalized deflation (GPCA-M).

For comparison, we also add the following sPCA techniques not based on loading deflation and that are described in the first paper of the series [5]. These will serve as a baseline to show the consequence of the introduction of artifacts by loading deflation:

1. PMD with scores orthogonalization [2] (PMD-O), instead of deflation.
2. The sPCA algorithm by Zou et al. [6] (SPCA-Z), which uses a simultaneous approach for model fitting.

5.2. Simulations

5.2.1. Orthogonal spectra

The orthogonal spectra are shown in Fig. 2, together with their approximation by the selected set of sPCA methods. The data set is generated from five mixtures of two compounds where the concentrations of the compounds (**T**) are represented at the top level of the figure and the pure spectra of the compounds (**P**) at the lower level of the figure.

No noise is introduced in the data, which is purely rank two. Loadings follow a shape of spectra, are intentionally non-negative and non-overlapping (thus orthogonal) and sparse. All sPCA methods can model this noise-free data set with perfect accuracy, which reflects an absence of rotational freedom.

5.2.2. Non-orthogonal spectra

The non-orthogonal spectra are shown in Fig. 3 with their approximation by the selected set of sPCA methods. The data set is generated from five mixtures of three compounds. There is some overlap (common variables/wavelengths) in the spectra of the compounds. Again, no noise is introduced in the data, which is purely rank three. Scores for the first component are not orthogonal to the other two, and they are all non-negative and sparse. Note that no sPCA method considered in this series can model this noise-free data set with perfect accuracy, even though data was simulated with a perfect sPCA model following Eq. (16).

5.2.3. Monte Carlo simulations

In order to generalize the results of the spectra examples, we randomly generated a set of 100 data sets in which the components can have a random number of non-zero loadings and the values of the scores and loadings are also randomly selected. Each data set $\mathbf{X}$ is generated with dimension 50×200 following a 5-component model with no noise. This means that the data is exactly rank five. The loading vectors have different degree of sparsity, contain both positive and negative values and are usually correlated because non zero elements can overlap. The scores can be non-orthogonal, are non-sparse and contain either positive or negative values.

5.2.4. Selection of metaparameters

All sparse algorithms considered require the setting of one or more metaparameters controlling the level of sparsity attained by the model. PMD models and SPCA-Z, which allow for the selection of the number of non-zeros elements, were set to meet the true number in the data generation. We set the GPCA models to meet, on average, the same level of sparseness as the other methods. In this way all methods can be fairly compared.

5.2.5. Accuracy statistics

Besides diagnostics in Section 4.4, which can be applied to real life cases, we can define additional statistics that can be applied on simulated data where the underlying data generation procedure is known. These are:

- AccL_A: The accuracy of identification of the sparse loadings, measured as the congruence between the true loadings and their estimates. This measure is akin to the absolute value of the Tucker's φ congruent coefficient [13,14], and takes values between 0 and 1.
- AccS_A: The accuracy of identification of the sparse scores, measured as the congruence between the true (known) scores and their estimates normalized to length 1. This measure is akin to the absolute value of the Tucker's congruent coefficient and varies between 0 and 1.

5.3. Experimental data

We consider 23 mixtures prepared by mixing 5 chemicals with varying concentrations according to a predetermined design (see Table 1 in the supplementary material of the original publication [15]). The chemicals were two peptides (valine-tyrosine-valine and valine-tyrosine-valine, a single amino acid (phenylalanine), a sugar (maltoheptaose) and an alcohol (propanol). The 23 mixtures and the five pure chemicals were prepared in aqueous phosphate buffer and acquired using an NMR spectrometer operating at a nominal frequency of 500 MHz. For details on sample preparation and NMR experiments we refer the reader to the original publication. Data were downloaded from www.models.life.ku.dk/joda/prototype. The original high resolution NMR data was binned to reduce dimensionality. The final data matrix of the mixtures has size 23×634 and the pure chemical compound matrix has size 5×634. Spectra are scaled to have maximum equal to 1 for ease of visualization. Analysis has been performed on the original bucketed values.

5.4. Software

All calculations have been performed in Matlab using the Matlab MEDA-toolbox [16] which is freely available at the address: github.com/josecamachop/MEDA-Toolbox. The VarA_A and angD_A statistics can be calculated using the function sparseart.m which takes as inputs the loadings and the scores calculated with any given sparse PCA algorithm.

6. Results

6.1. Simulation

We calculated these statistics for the sparse models fitted to the orthogonal spectra, the non-orthogonal spectra and the Monte Carlo simulations.

Table 1 shows the results for the orthogonal spectra. Since the example has two components, AngD_A, AccL_A and AccS_A, which are computed from the model part of the factorization, go from $A = 1$ to $A = 2$, while VarA_A, which is computed from the residuals, takes $A = 2$. The

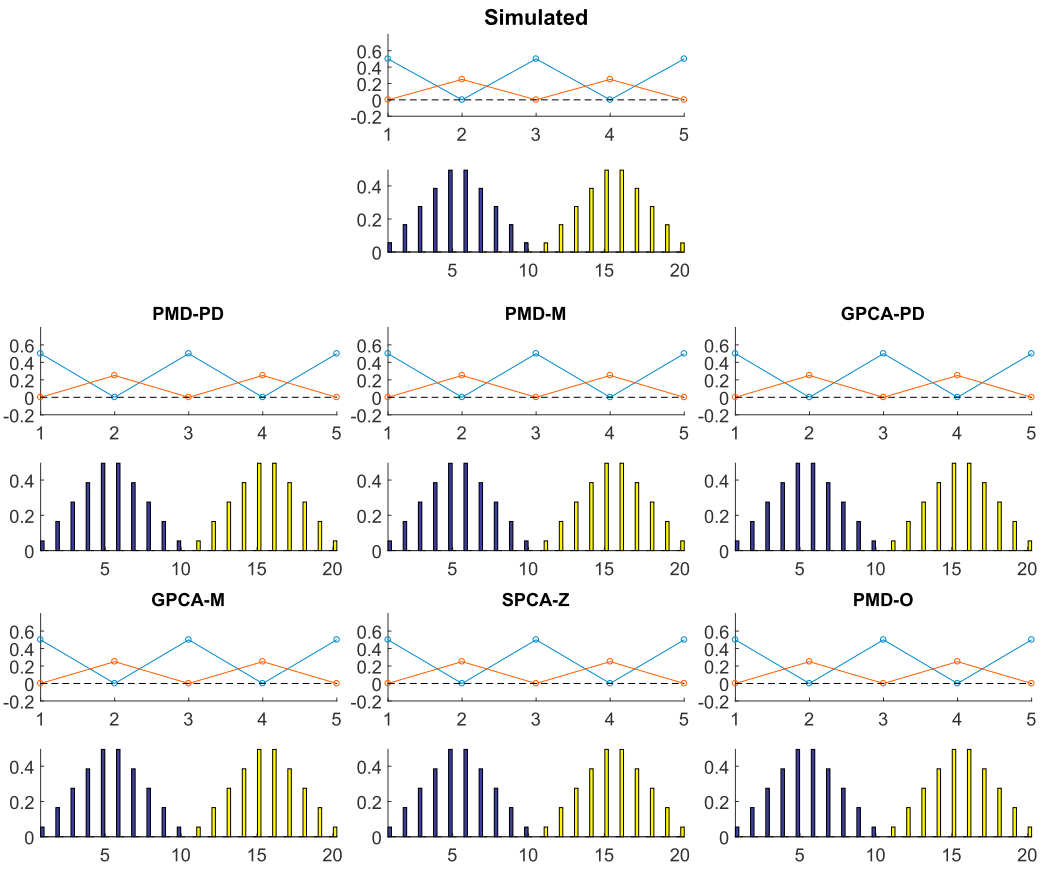

Fig. 2. Orthogonal spectra example: true data and selected sPCA methods. True and approximate concentrations (scores) of the five mixtures are represented at the top of the figures. Pure and approximate spectra (loadings) for twenty wavelengths are at the bottom of the figures. Please, refer to Ref. [5] for more details.

$AngD_A$ reflect all sPCA loadings are inside the row-space. Moreover, as expected for their orthogonality, no artifacts ($VarA_2$) are found in the components, and model present a perfect fit (RSS) and accuracy ($AccL_A$ and $AccS_A$).

Table 2 shows the results for the non-orthogonal spectra. Since the example has three components, $AngD_A$, $AccL_A$ and $AccS_A$ go from $A = 1$ to $A = 3$, while $VarA_A$ goes from $A = 2$ to $A = 3$. According to the $AngD_A$, all sPCA loadings are outside the row-space, with a maximum angle above 45° degrees of deviation. Generally speaking, this is an expected behavior of sPCA models, since they use some form of penalties or constraints. However, in this case the data is noise-free and follows a perfect sparse component model. Yet, the first component gets outside the data row-space. The first component is shown in Fig. 3 and it is simulated to have non-zero loadings in the first 10 variables. However, none of the sPCA algorithms select the correct 10 variables in the first component, probably as a result of rotational ambiguity and/or algorithmic inaccuracies. This has more dramatic implications in deflation-based methods, because residuals get outside the data row-space and artifacts are generated within the following components. Artifacts can corrupt up to 50% ($VarA_2$) of the variance of components fitted after the first sparse loading was deflated, and close to 80% ($VarA_3$) after the first two loadings. This does not affect sPCA algorithms that do not use deflation: while we can still compute $VarA_A$ for them, their value remains low and does not reflect contamination with artifacts. We can generally see a higher fit (lower RSS) in deflation methods and a direct relationship between the generation of artifacts $VarA_A$ and the accuracy of the subsequent loading ($AccL_A$). Moreover, for the deflation-based approaches, accuracy tends to deteriorate with the order of the component, something that does not happen for SPCA-Z or PMD-O. The accuracy of the scores ($AccS_A$) remains similar among the approaches, except for the third component in deflation-based PMD. This may be the consequence of the loss of accuracy in the third loading. As a whole, we can see a direct relationship between the artifacts and the loadings accuracy (poorer for deflation-based approaches) which interestingly comes with a higher fit. The results in Table 2 reflect very well what is observed in Fig. 3.

Fig. 4 generalizes the previous results using the Monte Carlo experiment. Recall that this example has five sparse components. Fig. 4a shows that after the five are obtained, SPCA-Z and PMD-O remain close to the row-space, while the deflation-based approaches get almost completely outside. We see that the latter tend to leave the row-space increasingly with the number of components, since the angle ($AngD_A$) is monotonous increasing. This is consistent with what observed in the non-orthogonal spectra example. The percentage of artifacts induced by deflation-based approaches is worrisome, reaching occasionally the 100% for the fourth and fifth components. Accuracy in loadings and scores is much worse in deflation-based approaches, especially in higher order components as a result of the high percentage of artifacts. Therefore, for noise-free random data and optimal selection of metaparameters, we can conclude that non-deflation approaches outperform those based on the deflation. Interestingly, again, this improvement does not correlate with fit: sPCA variants with high fit are less accurate. This means that the popular approach to assess the goodness of sPCA models based on captured variance [6,11] is not generally valid.

6.2. Real example

To show the problems and limitations arising when using deflation based approaches to extract higher order sparse components we analyzed an experimental data **X** containing 23 mixtures of 5 different pure chemical compounds. The NMR spectra of the mixtures are given in Fig. 5A, and the spectra for the pure compounds are shown in Fig. 5B.

The data matrix **X** can be then decomposed using an sPCA model. We

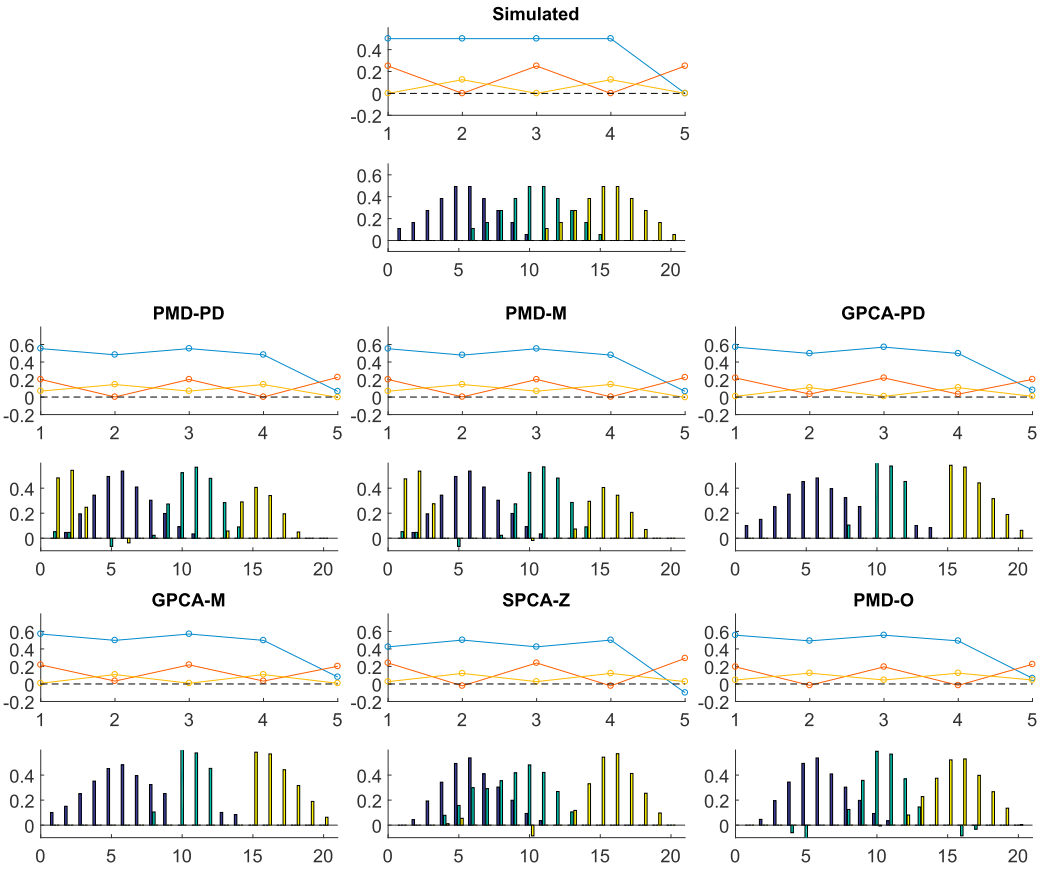

Fig. 3. Non-orthogonal spectra example: true data and selected sPCA methods. True and approximate concentrations (scores) of the five mixtures are represented at the top of the figures. Pure and approximate spectra (loadings) for twenty wavelengths are at the bottom of the figures. Please, refer to Refs. [5] for more details.

Table 1
Performance statistics of sPCA variants for the orthogonal spectra.

	PMD-PD	PMD-M	GPCA-PD	GPCA-M	SPCA-Z	PMD-O
AngD$_1$	0	0	0	0	0	0
AngD$_2$	0	0	0	0	0	0
VarA$_2$	0	0	0	0	0	0
RSS	0	0	0	0	0	0
AccL$_1$	1	1	1	1	1	1
AccL$_2$	1	1	1	1	1	1
AccS$_1$	1	1	1	1	1	1
AccS$_2$	1	1	1	1	1	1

Table 2
Performance statistics of sPCA variants for the non-orthogonal spectra. * Statistic VarA$_A$ is non-applicable to (meaningless for) sPCA methods not based on deflation, but it can still be computed.

	PMD-PD	PMD-M	GPCA-PD	GPCA-M	SPCA-Z	PMD-O
AngD$_1$	10.6	9.1	9.6	9.6	9.3	9.1
AngD$_2$	19.1	15.0	21.5	21.5	12.4	15.2
AngD$_3$	46.9	36.7	27.5	27.5	11.9	4.0
VarA$_2$	20.4	17.2	51.1	51.1	NA (13.4)*	NA (7.4)*
VarA$_3$	77.5	68.1	37.9	37.9	NA (12.3)*	NA (2.3)*
RSS	0.020	0.015	0.019	0.019	0.042	0.032
AccL$_1$	0.975	0.980	0.983	0.983	0.979	0.980
AccL$_2$	0.915	0.931	0.818	0.818	0.916	0.925
AccL$_3$	0.478	0.602	0.859	0.859	0.953	0.987
AccS$_1$	0.997	0.996	0.995	0.995	0.991	0.996
AccS$_2$	0.997	0.999	0.991	0.991	0.993	0.996
AccS$_3$	0.884	0.907	0.996	0.996	0.965	0.913

used the PMD algorithm [2] with projection deflation (PMD-PD) to extract sparse loadings under the assumption that each sparse loading vector should represent a different chemical compound, *i.e.* the spectra of the NMR. Note that the pure spectra are overlapping. Fitting a sPCA model with the PMD requires to determine *a priori* the level of sparseness, *i.e.* the number of non-zero loadings to retain for each component. In real-life applications the level of sparsity is not known since the data generation mechanism is also not known, especially in the case of complex biological data. In this case, however, the data structure is known and the most convenient sparsity can be inferred from the pure spectra. Basically, we first sorted the pure spectra in descending order of variance and then computed the number of variables with values larger than 0.05 for each pure spectra assigning it as the sparsity level for the first, second, etc component. The number of non-zero loadings for each component was set to 38, 31, 28, 37, and 12.

Since the data is a pure rank-five system, we fit a sparse model with five components. The sparse loadings are given in Fig. 5C, side to side to the pure spectra (panel B). The first and the fifth components agree, albeit not perfectly, with the fourth and the fifth pure spectra as shown in Fig. 5D, but for the other components there is little to no agreement with the pure spectra. Also, there is not a good recovery of the true concentration matrix, as shown in Fig. 5E with the AccS$_A$ statistics. These problems are further summarized by the two statistics AngD$_A$ and varA$_A$, which are given in Fig. 5F: components one and five show reasonably low statistics, which are higher in the case of the other components. The value of varA$_1$ is by definition 0. The value of varA$_5$ is reasonably low, given 4 components have been deflated before.

For real life applications we suggest to report the two statistics AngD$_A$ and varA$_A$ for any deflation-based sPCA model and a cautious interpretation of the sparse components whenever those statistics are excessively large.

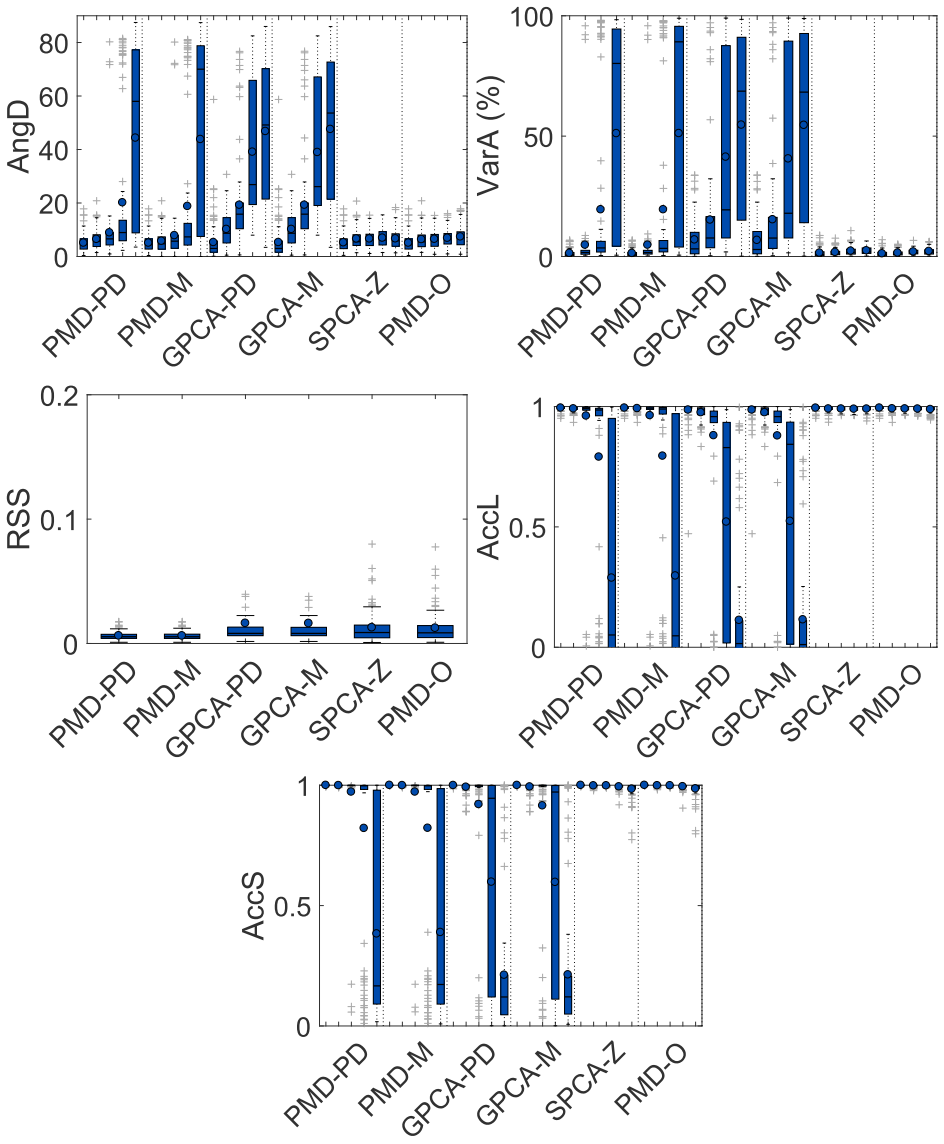

Fig. 4. Box Plots of the statistics for the Monte Carlo simulation.

7. Conclusion

In Part I of this series, we described limitations of several state-of-the-art Sparse Principal Component Analysis (sPCA) algorithms when modeling noise-free data, simulated following an exact sPCA model. In this second paper, we show both theoretically and numerically that loading deflation can lead to the inclusion of artifacts or unexpected variability from the second component onwards. This may result in a wrong interpretation of the sPCA model. We also provide diagnostics that can be used to detect the problem in practical applications: the angle with the data row-space ($AngD_A$) and the percentage of artifacts (Var_A). Furthermore, we propose a procedure to safely interpret sPCA models with loading deflation using these statistics. We suggest to report the statistics together with the sparse models in the literature where interpretation is the modeling goal.

The described problems do not affect sparse loadings that are orthogonal. One sub-class of orthogonal loadings are sparse loadings that do not overlap (do not share common variables). It is useful to know in practical applications that sPCA models with non-overlapping loadings can be safely identified with deflation based algorithms.

Finally, as we showed in the examples, there is an interplay between rotational ambiguity in sPCA models and the introduction of spurious variance due to deflation. Future work is needed to establish how rotational ambiguity can affect the interpretation of sPCA models in general.

CRediT author contribution statement

J. Camacho: All authors were equally involved in the research and preparation of the paper. **A.K. Smilde:** All authors were equally involved in the research and preparation of the paper. **E. Saccenti:** All authors were equally involved in the research and preparation of the paper, All authors were equally involved in the research and preparation of the paper. **Rasmus Bro:** All authors were equally involved in the research and preparation of the paper.

Declaration of competing interest

The authors declare that they have no known competing financial

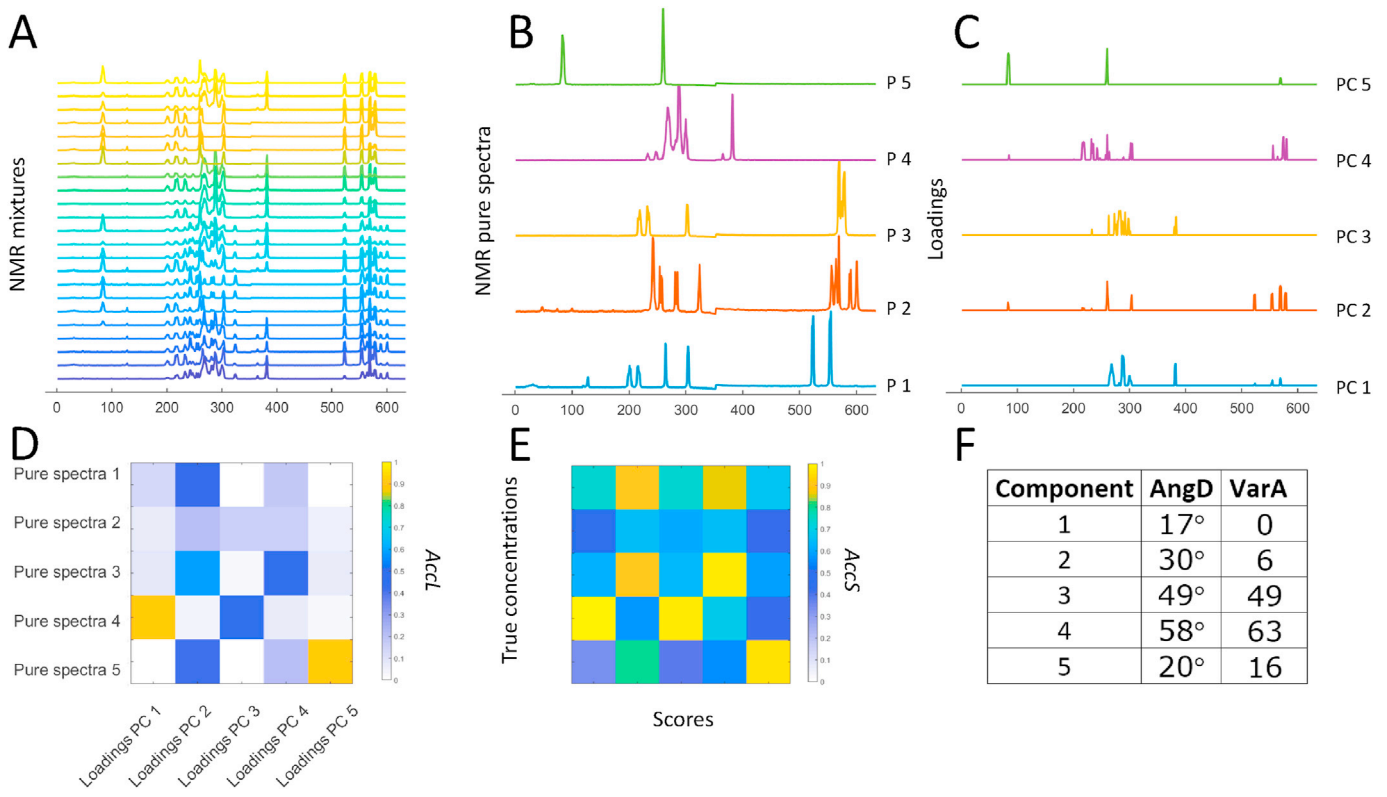

Fig. 5. Sparse PCA analysis of a data sets containing 23 mixtures of 5 pure chemical compounds. The sparse PCA model is obtained using the PMD algorithm [2]. A) NMR spectra of the 23 chemical mixtures. B) NMR spectra of the 5 pure chemical mixtures (P1: tryptophane-glycine, P2: phenylalanine, P3: maltoheptaose, P4: valine-tyrosine-valine, P5: propanol). C) Loading of the 5 sparse components. D) Agreement (*AccL*) between the pure spectra and the sparse loadings. E) Agreement (*AccS*) between the true concentration matrix and the model scores. F) Summary of AngD_A and Var_A statistics summarizing departure from the data row-space and variance due to artifacts. The pure spectra are ordered following the congruence of the loadings of the sparse PCA model given in panel D. x-axis for panel A, B and C indicates the NMR variable index. Y-axis is arbitrary units.

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is partly supported by the Spanish Ministry of Economy and Competitiveness and ERDF (European Regional Development Fund) funds through project TIN2017-83494-R and the "Plan Propio de la Universidad de Granada" and by The Netherlands Organisation for Health Research and Development (ZonMW) through the PERMIT project (Personalized Medicine in Infections: from Systems Biomedicine and Immunometabolism to Precision Diagnosis and Stratification Permitting Individualized Therapies, project number 456008002) under the PerMed Joint Transnational call JTC 2018 (Research projects on personalized medicine - smart combination of pre-clinical and clinical research with data and ICT solutions).

Appendix A. Deflation in PMD

Reference [2] describes the sPCA particularization of the more general PMD as an algorithm that follows Eq. (10) and where $\hat{\mathbf{p}}_A^P$ is obtained by soft-thresholding. Given $\hat{\mathbf{p}}_A^P$, Eq. (10) can be re-stated as:

$$\hat{\mathbf{u}}_A^P = argmax_{\mathbf{u}_A}\ \mathbf{u}_A^{\mathrm{T}}\mathbf{X}_{A-1}\hat{\mathbf{p}}_A^P, \quad s.t. \quad \|\mathbf{u}_A\|_2^2 \leq 1. \tag{A.1}$$

The least squares solution is given by the Lagrange method, defining function:

$$F = \mathbf{u}_A^{\mathrm{T}}\mathbf{X}_{A-1}\hat{\mathbf{p}}_A^P + \lambda(\mathbf{u}_A^{\mathrm{T}}\mathbf{u}_A - 1), \tag{A.2}$$

and solving for $\frac{\partial F}{\partial \mathbf{u}_A} = 0$ and $\frac{\partial F}{\partial \lambda} = 0$. The former leads to the expression:

$$\hat{\mathbf{u}}_A^P = \frac{\mathbf{X}_{A-1}\hat{\mathbf{p}}_A^P}{\lambda}, \tag{A.3}$$

while the latter simply sets $\hat{\mathbf{u}}_A^P$ to unit length, that is, sets $\lambda = \|\mathbf{X}_{A-1}\hat{\mathbf{p}}_A^P\|_2$:

$$\hat{\mathbf{u}}_A^P = \frac{\mathbf{X}_{A-1}\hat{\mathbf{p}}_A^P}{\|\mathbf{X}_{A-1}\hat{\mathbf{p}}_A^P\|_2}. \tag{A.4}$$

Subsequently, the least squares solution, given $\hat{\mathbf{p}}_A^P$ and $\hat{\mathbf{u}}_A^P$, of loss:

$$\hat{d}_A^P = argmax_{d_A} \ \|\mathbf{X}_{A-1} - d_A\hat{\mathbf{u}}_A^P(\hat{\mathbf{p}}_A^P)^{\mathrm{T}}\|_2^2 \tag{A.5}$$

is derived from:

$$\frac{\partial\left\|\mathbf{X}_{A-1} - d_A\hat{\mathbf{u}}_A^P(\hat{\mathbf{p}}_A^P)^{\mathrm{T}}\right\|_2^2}{\partial d_A} = 2\cdot(d_A\cdot\hat{\mathbf{p}}_A^P(\hat{\mathbf{u}}_A^P)^{\mathrm{T}} - \mathbf{X}_{A-1}^{\mathrm{T}})\hat{\mathbf{u}}_A^P(\hat{\mathbf{p}}_A^P)^{\mathrm{T}} = 0. \tag{A.6}$$

yielding Eq. (11).

Appendix B. Sparse vectors and data column/row-space

Appendix B.1. Proofs of Case 1: Low-dimensional case $M < N$

Assuming data with noise, it holds that $R(\mathbf{X}_1^{\mathrm{T}}) = \mathbb{R}^M$. Since the rows of $\mathbf{X}_2$ are M-dimensional it holds that $R(\mathbf{X}_2^{\mathrm{T}})\subseteq\mathbb{R}^M$ and thus $R(\mathbf{X}_2^{\mathrm{T}})\subseteq R(\mathbf{X}_1^{\mathrm{T}})$. The row-wise deflation can be written as $\mathbf{X}_2 = \mathbf{X}_1(\mathbf{I} - \mathbf{p}\mathbf{p}^{\mathrm{T}})$ assuming that $\|\mathbf{p}\|_2 = 1$, for convenience and without lack of generality. Hence, $\mathbf{X}_1^{\mathrm{T}} = \mathbf{X}_2^{\mathrm{T}} + \mathbf{p}\mathbf{p}^T\mathbf{X}_1^T$. Both $R(\mathbf{p})$ and $R(\mathbf{X}_2^{\mathrm{T}})$ are in $R(\mathbf{X}_1^{\mathrm{T}})$ and $R(\mathbf{p})\cap R(\mathbf{X}_2^{\mathrm{T}}) = 0$, thus $R(\mathbf{X}_1^{\mathrm{T}}) = R(\mathbf{X}_2^{\mathrm{T}})\oplus R(\mathbf{p}\mathbf{p}^{\mathrm{T}})$, where the symbol $\oplus$ indicates the direct sum. Consequently, $\dim(R(\mathbf{X}_1^{\mathrm{T}})) = \dim(R(\mathbf{X}_2^{\mathrm{T}})) + \dim(R(\mathbf{p}\mathbf{p}^{\mathrm{T}}))$ or $M = (M-1)+1$ since $\dim(R(\mathbf{p}\mathbf{p}^{\mathrm{T}})) = 1$. Thus, the row-deflation 'peels of' dimensions of $\mathbf{X}_1$ one by one.

Appendix B.2. Proofs of Case 2: High-dimensional case $N < M$

There are two possibilities: $R(\mathbf{p})\subseteq R(\mathbf{X}_1^{\mathrm{T}})$ (case 2a) or $R(\mathbf{p})\nsubseteq R(\mathbf{X}_1^{\mathrm{T}})$ (case 2b). The analysis of case 2a runs along the same lines as that of case 1 and will not be repeated. Case 2b goes as follows. The starting equation is again $\mathbf{X}_2 = \mathbf{X}_1(\mathbf{I} - \mathbf{p}\mathbf{p}^{\mathrm{T}})$, but since $R(\mathbf{p})\nsubseteq R(\mathbf{X}_1^{\mathrm{T}})$ it holds now that $R(\mathbf{X}_2^{\mathrm{T}})\nsubseteq R(\mathbf{X}_1^{\mathrm{T}})$. To find the discrepancy between $R(\mathbf{X}_2^{\mathrm{T}})$ and $R(\mathbf{X}_1^{\mathrm{T}})$, the residuals $\mathbf{E}$ of projecting $R(\mathbf{X}_2^{\mathrm{T}})$ onto $R(\mathbf{X}_1^{\mathrm{T}})$ can be studied and these are $\mathbf{E} = (\mathbf{I} - \mathbf{X}_1^{\mathrm{T}}(\mathbf{X}_1^{\mathrm{T}})^{+})\mathbf{X}_2^{\mathrm{T}}$ (where the superscript '+' indicates the Moore-Penrose (MP) inverse. Substituting in $\mathbf{X}_2 = \mathbf{X}_1(\mathbf{I} - \mathbf{p}\mathbf{p}^{\mathrm{T}})$ in this equation and working it out using the properties of the MP inverse shows that $\mathbf{E} = (\mathbf{X}_1^{\mathrm{T}}(\mathbf{X}_1^{\mathrm{T}})^{+} - \mathbf{I})\mathbf{p}\mathbf{p}^{\mathrm{T}}\mathbf{X}_1^{\mathrm{T}}$ which can be written schematically as $\mathbf{E} = \mathbf{X}_1^{\mathrm{T}}\mathbf{A} - \mathbf{p}\mathbf{b}^T$. This equation shows cleary what happens in the two different situations: if $R(\mathbf{p})\subseteq R(\mathbf{X}_1^{\mathrm{T}})$ then $R(\mathbf{E})\subseteq R(\mathbf{X}_1^{\mathrm{T}})$ and if $R(\mathbf{p})\nsubseteq R(\mathbf{X}_1^{\mathrm{T}})$ then $R(\mathbf{E})\nsubseteq R(\mathbf{X}_1^{\mathrm{T}})$ and the more $\mathbf{p}$ is outside $R(\mathbf{X}_1^{\mathrm{T}})$, the more $R(\mathbf{E})$ will deviate from $R(\mathbf{X}_1^{\mathrm{T}})$.

Appendix B.3. High-dimensional case: consecutive components

Let us study the subspaces where consecutive loadings lie in the high-dimensional case. Take $\mathbf{p}_1$ and $\mathbf{p}_2$, fitted from $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Generally speaking, we can write $\mathbf{p}_1 = \mathbf{p}_1^0 + \mathbf{p}_1^{\perp 0}$, so that $\mathbf{p}_1^0$ is the projection of $\mathbf{p}_1$ onto $R(\mathbf{X}_1^{\mathrm{T}})$ and $\mathbf{p}_1^{\perp 0}$ the orthogonal part. Then, it holds:

- $R(\mathbf{p}_1^0)\subseteq R(\mathbf{X}_1^{\mathrm{T}})$
- $R(\mathbf{p}_1^{\perp 0})\perp R(\mathbf{X}_1^{\mathrm{T}})$
- $(\mathbf{p}_1^0)^{\mathrm{T}}\mathbf{p}_1^{\perp 0} = 0$

We can interpret that:

- $\mathbf{p}_1^0$: is the part of $\mathbf{p}_1$ that can be trace back to the data in $\mathbf{X}_1$.
- $\mathbf{p}_1^{\perp 0}$: is the part of $\mathbf{p}_1$ that actually makes it sparse.

In the same way, we can write $\mathbf{p}_2 = \mathbf{p}_2^1 + \mathbf{p}_2^{\perp 1}$, so that $\mathbf{p}_2^1$ is the projection of $\mathbf{p}_2$ in $R(\mathbf{X}_2^{\mathrm{T}})$ and $\mathbf{p}_2^{\perp 1}$ the orthogonal part. Given that $R(\mathbf{X}_2^{\mathrm{T}})\nsubseteq R(\mathbf{X}_1^{\mathrm{T}})$ in the high-dimensional case, we can further divide $\mathbf{p}_2^1$ into its projection onto $R(\mathbf{X}_1^{\mathrm{T}})$ and the orthogonal part: $\mathbf{p}_2^1 = \mathbf{p}_2^{1,0} + \mathbf{p}_2^{1,\perp 0}$, so that at the end we have: $\mathbf{p}_2 = \mathbf{p}_2^{1,0} + \mathbf{p}_2^{1,\perp 0} + \mathbf{p}_2^{\perp 1}$, where:

- $\mathbf{p}_2^{1,0}$: is the part of $\mathbf{p}_2$ that can be traced back to $\mathbf{X}_2$ and to the data in $\mathbf{X}_1$.
- $\mathbf{p}_2^{1,\perp 0}$: is the part of $\mathbf{p}_2$ that can be traced back to $\mathbf{X}_2$ but not to $\mathbf{X}_1$.
- $\mathbf{p}_2^{\perp 1}$: is the part of $\mathbf{p}_2$ that actually makes it sparse.

Equations (14) and (15) are intended to quantify $\mathbf{p}_2^{1,\perp 0}$ or, generally speaking, $\mathbf{p}_{A+1}^{A,\perp 0}$, as described in the main body of the manuscript.

References

[1] Mackey Lester, Deflation methods for sparse PCA, NIPS (News Physiol. Sci.) (2008) 1–8.

[2] M. Witten Daniela, Tibshirani Robert, Hastie Trevor, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (2009) 515–534.

[3] Hastie Trevor, Tibshirani Robert, Wainwright Martin, Statistical Learning with Sparsity: the Lasso and Generalizations, Chapman & Hall/CRC, 2015.

[4] Hui Zou, Lingzhou Xue, A selective overview of sparse principal component analysis, Proc. IEEE 106 (2018) 1311–1320.

[5] J. Camacho, A.K. Smilde, E. Saccenti, J.A. Westerhuis, All sparse PCA models are wrong, but some are useful. Part I: computation of scores, residuals and explained variance, Chemometr. Intell. Lab. Syst. 196 (2020) 103907.

[6] Hui Zou, Hastie Trevor, Tibshirani Robert, Sparse principal component analysis, J. Comput. Graph Stat. 15 (2006) 265–286.

[7] Camacho José, A. Rodríguez-Gómez Rafael, Saccenti Edoardo, Group-wise principal component analysis for exploratory data analysis, J. Comput. Graph Stat. 26 (2017) 501–512.

[8] Richtárik Peter, Takác Martin, Ahipasaoglu Selin Damla, Alternating Maximization: Unifying Framework for 8 Sparse PCA Formulations and Efficient Parallel Codes *CoRR*, 2012 abs/1212.4137.

[9] I.T. Jolliffe, Principal Component Analysis, EEUU: Springer Verlag Inc, 2002.

[10] A. Westerhuis Johan, K. Smilde Age, Deflation in multiblock PLS, J. Chemom. 15 (2001) 485–493.

[11] I.T. Jolliffe, N.T. Trendafilov, M. Uddin, A modified principal component technique based on the LASSO, J. Comput. Graph Stat. 12 (3) (2003) 531–547.

[12] Camacho Jose, Evrim Acar, A. Rasmussen Morten, Bro Rasmus, Cross-product penalized component analysis (X-CAN), Chemometr. Intell. Lab. Syst. 203 (2020) 104038.

[13] R. Tucker Ledyard, A Method for Synthesis of Factor Analysis Studies Tech. Rep, Educational Testing Service, Princeton Nj, 1951.

[14] Lorenzo-Seva Urbano, M.F. Ten Berge Jos, Tucker's congruence coefficient as a meaningful index of factor similarity, Methodology 2 (2006) 57–64.

[15] Evrim Acar, E. Papalexakis Evangelos, Gürdeniz Gözde, et al., Structure-revealing data fusion, BMC Bioinf. 15 (2014) 1–17.

[16] Camacho José, Pérez-Villegas Alejandro, A. Rodríguez-Gómez Rafael, Jiménez-manas elena. Multivariate exploratory data analysis (MEDA) toolbox for Matlab, Chemometr. Intell. Lab. Syst. 143 (2015) 49–57.