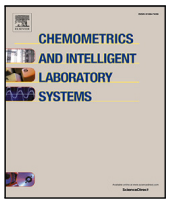
Chemometrics and Intelligent Laboratory Systems 266 (2025) 105498

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 III: Model interpretation

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

[a] Computational Data Science Laboratory, Research Centre for Information and Communication Technologies (CITIC-UGR), 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:
Exploratory data analysis
Model interpretation
Sparse principal component analysis
Sparsity

ABSTRACT

Sparse Principal Component Analysis (sPCA) is a popular matrix factorization that combines variance maximization and sparsity with the ultimate goal of improving data interpretation. In this series of papers we show that the factorization with sPCA can be complex to interpret even when confronted with simple data. In the first paper in this series, we demonstrated that sPCA models have limitations with respect to factorizing sparse and noise-free data accurately when loadings are overlapping. In the second paper, we showed that sPCA algorithms based on deflation can generate artifacts in high order components. We also show that scores orthogonalization and the incorporation of orthonormal loadings are suitable means to avoid large artifacts. Both approaches constrain the set of possible sPCA solutions in a very similar but poorly understood way. In particular, we study in this paper the sPCA solution by Zou et al., which according to our results represent the best sPCA algorithm of those considered in the series. Here, we provide new derivations on the model equations, the computation and interpretation of the model parameters and the selection of metaparemeters in practical cases, making sPCA an even more powerful data modeling tool.

1. Introduction

Model interpretation is a critical step in sparse principal component analysis (sPCA) [1–3]. In this series of papers "All sparse PCA models are wrong, but some are useful", we have identified and discussed modeling problems and interpretation challenges affecting the most popular implementations of sPCA.

In the first paper of the series [4], we illustrated the limitations of sPCA models for factorizing noise-free and exactly sparse data (with many exact zero values) when the true loadings overlap. It was shown that even in this simple case, there are severe precautions that need to be taken when interpreting the models. We also observed that the most commonly-used implementations of sPCA either underestimate or overestimate the correlation between pairs of loadings and scores; and we proposed adjustments for the estimation of the scores and the explained variance which, albeit not fully correcting for these problems, may significantly increase the quality of model estimation.

In the second paper [5], we focused on the use of deflation, a popular algorithmic approach within sPCA and many other multivariate algorithms. Sparse loadings, like any form of constrained loadings, may be outside the rowspace of the data and we showed that, as a consequence, deflation can lead to the inclusion of artifacts (fake patterns not found in the data) in the estimated loadings from the second component onwards. The inclusion of spurious information in the sparse components may result in loss of accuracy and lead to wrong interpretation of the sPCA model.

This type of problem does not manifest when the sparse loadings are orthogonal, which in particular is the case when there is no overlap, *i.e.,* when loadings of different components do not share common (non-zero) variables. Generally speaking, we found that deflation-based sPCA algorithms are not an adequate choice when the data conforms to non-orthogonal sparse loadings structures.

We provided two diagnostics that can be used to detect and quantify the inclusion of spurious information in practical applications: the percentage of artifacts and the angle with the data rowspace [5]. When reporting deflation-based sPCA results in the literature, we suggested also reporting the diagnostics, to gauge the appropriateness of fitting a sparse model to the data. With this suggestion we imply that models with large values for these diagnostics are probably not adequate for data interpretation.

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

https://doi.org/10.1016/j.chemolab.2025.105498
Received 10 December 2024; Received in revised form 12 June 2025; Accepted 25 July 2025
Available online 12 August 2025

In reviewing sPCA algorithms, we identified two algorithmic approaches that can control the departure of sPCA from the data rowspace, namely: the use of orthonormal loadings in the sPCA algorithm [6,7], and the version with scores orthogonalization of the Penalized Matrix Decomposition (PMD) [2]. We showed how the use of these two approaches together with our corrections can outperform deflation-based sPCA in terms of (sparse) modeling and data fitting. Both approaches constrain the set of possible sPCA solutions in a very similar but poorly understood way, since their assumptions on the data generation process are quite different [8], which can actually lead to an incorrect interpretation of the models.

Underlying the use and interpretation of sparse component models, and of sparse PCA modeling in particular, is the idea that a sparse model is simpler to interpret because the number of selected variables is smaller, and often much smaller, than the original number of variables. However, sparsity, interpretation, and interpretability are actually distinct concepts.

Interpretability is a property of a (data analysis) method related to its capacity to produce and present results that can be interpreted within the body of actual knowledge. The interpretation of the model, i.e., the interpretation of the results, is the interplay between the presented results by a method and the domain expert with the goal of inferring meaning from the results, thereby enlarging the understanding of the studied system.

The reduction in the number of variables, i.e., the sparsity of the model, is taken as a proxy for enhanced interpretability, since it is often assumed that only a subset of measured variables is related to the problem being studied [9], as measured variables may include informative, redundant and non-informative variables. This can be defined as the "sparse truth" assumption.

A legitimate question to ask is whether sparsity is the true reality for a given studied system (either biological, chemical, physical, or other) and whether this is visible in the data. We can use a sparse model under the expectation that the underlying truth is inherently sparse and under the assumption that it is able to capture this type of sparseness. Yet, a more practical and widely applicable approach is to assume that the underlying truth and subsequent data is not necessarily sparse, and that the sparse model only approximates this truth and is, hopefully, more interpretable than a non-sparse model.

In the present paper, that conclude this series, we address the problem of model interpretation: upon characterization of the algebraic properties of the sPCA solution, we investigate which model parameters carry more or more accurate information about the underlying truth and how they should be interpreted. We focus on the particular sPCA solution by Zou et al. [6], which outperforms other solutions in explained variance [7] and interpretation [2]. We also revisit the results of Guerra-Urzola et al. [10], which showed limited performance of the sPCA solution by Zou et al.under certain conditions. Our analyses illustrate that sPCA can indeed yield good performance when following our strategy even in those conditions.

The rest of the paper is organized as follows. Section 2 reviews the sPCA algorithm proposed by Zou et al. [6] and motivates our choice of focusing on this method based on the incorporation of orthonormal loadings. Section 3 provides a derivation of the underlying model by Zou et al., and deals with the definition and interpretation of the scores, loadings and explained variance. Section 4 presents the analysis of a benchmark experimental data set and revisits some of the results by Guerra-Urzola et al. Section 5 provides a general discussion of the series and concluding remarks.

2. Sparse PCA algorithms

We detail below the sPCA algorithm by Zou et al. [6], which implements a simultaneous[1] approach to extract multiple sparse components making use of auxiliary non-sparse loadings. In Appendix, we briefly describe the sequential version of the Zou et al. algorithm [7], and a sequential sPCA variant based on scores orthogonalization [2]. All these methods showed superior modeling capabilities when compared to deflation-based approaches in the second paper of the series [5], but the sequential version generally captures less variance than the sPCA algorithm by Zou et al.and the orthogonalization algorithm [2] is more difficult to interpret.

The sPCA algorithm by Zou et al.(in the following we will refer to this approach as SPCA-*Z*, with capital 'S' and 'Z' to differentiate it from the general sPCA approach) makes use of non-sparse orthonormal loadings to extract sparse components in a simultaneous fashion. The SPCA-*Z* model is based on the reformulation of the problem of finding an sPCA solution as a regularized regression problem with a criterion close to the (naive) elastic net [12], which is a combination of the lasso and the ridge constraints. To fit an sPCA model with A components to an $N \times J$ (observations $\times$ variables) data matrix $\mathbf{X}$, the optimization problem is given by

$$\underset{\mathbf{P},\mathbf{Q}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{X}\mathbf{P}\mathbf{Q}^{\mathrm{T}}\|_F^2 + \lambda_2 \sum_{a=1}^{A} \|\mathbf{p}_a\|_2^2 + \lambda_1 \sum_{a=1}^{A} \|\mathbf{p}_a\|_1 \quad s.t. \quad \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}, \quad (1)$$

where $\mathbf{P}$ and $\mathbf{Q}$ are the $J \times A$ matrices of sparse and orthonormal loadings, respectively, $\mathbf{p}_a$ is the ath column vector ($J \times 1$) in $\mathbf{P}$, $\mathbf{I}$ is the $A \times A$ identity matrix and $\| * \|_F$, $\| * \|_2$ and $\| * \|_1$ are the Frobenius norm of a matrix, the 2-norm and the 1-norm of a vector, respectively. The sparsity of the solution depends on the relative importance provided to these norms by the parameters λ_2 and λ_1, respectively. We maintain the notation of the sparse loadings ($\mathbf{P}$) to be consistent with our previous papers of the series, and we will refer to those as sparse weights, following Park et al. [8].

The numerical solution proposed for Eq. (1) is a biconvex optimization where sparse weights and orthonormal loadings are obtained using an alternating approach, which guarantees convergence to a local minimum. In the first step, we use the equivalence $\operatorname{argmin}_{\mathbf{P}} \|\mathbf{X} - \mathbf{X}\mathbf{P}\mathbf{Q}^{\mathrm{T}}\|_F^2 \equiv \operatorname{argmin}_{\mathbf{P}} \|\mathbf{X}\mathbf{Q} - \mathbf{X}\mathbf{P}\|_F^2$, which holds for $\mathbf{P}$ column-wise full-rank and comes from imposing the constraint $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$. This equivalence allows to solve $\mathbf{P}$ for fixed $\mathbf{Q}$ in Eq. (1) as:

$$\underset{\mathbf{P}}{\operatorname{argmin}} \|\mathbf{X}\mathbf{Q} - \mathbf{X}\mathbf{P}\|_F^2 + \lambda_2 \sum_{a=1}^{A} \|\mathbf{p}_a\|_2^2 + \lambda_1 \sum_{a=1}^{A} \|\mathbf{p}_a\|_1, \quad (2)$$

using A independent elastic net regressions for which the response is $\mathbf{X}\mathbf{q}_a$ and the corresponding sparse vectors $\mathbf{p}_a$ are estimated.

In the second step, the orthonormal loadings in $\mathbf{Q}$ are found for fixed $\mathbf{P}$ as the solution to the Procrustes problem:

$$\underset{\mathbf{Q}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{X}\mathbf{P}\mathbf{Q}^{\mathrm{T}}\|_F^2 \quad s.t. \quad \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}, \quad (3)$$

which is the polar decomposition:

$$\mathbf{Q} = \mathbf{U}\mathbf{V}^{\mathrm{T}}, \quad (4)$$

where $\mathbf{U}$ and $\mathbf{V}$ are the $J \times A$ and $A \times A$ matrices with the left and right singular vectors, respectively, from the (truncated) Singular Value Decomposition of the $J \times A$ matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}, \quad (5)$$

where $\mathbf{S}$ is $A \times A$. The two steps are iterated until convergence.

[1] The term simultaneous here means that the model fits all components at once, rather than in a sequential fashion, where one component is fitted at a time followed by deflation or orthogonalization. In other areas [11], a simultaneous approach is referred to as a block approach.

3. Interpretation of the solution of the SPCA-Z algorithm

Concerning the original presentation of the SPCA-Z algorithm, we note the following four interrelated issues:

1. The scores are computed in such a way that the resulting estimated variances are inaccurate, as we have previously shown [4].
2. Only the sparse weights **P** are considered for the interpretation of case studies [6,13].
3. The orthonormal loadings **Q** are used as a mere computational aid for model fitting and later discarded, leading to a loss of information which is fundamental for the interpretation of the resulting sparse model.
4. As a consequence, the model used for interpretation is inconsistent with the model underlying the fitting procedure, as we will derive mathematically later on.

By addressing points 1 to 4 in this section, we will show how the interpretability of the sPCA solution obtained from the SPCA-Z algorithm can be enhanced by taking advantage of a careful algebraic characterization.

3.1. The model underlying SPCA-Z

We derive the model underlying the solution at convergence of the SPCA-Z algorithm, for a model fitted with A components to an $N \times J$ data matrix **X** and resulting in sparse weights **P** and orthonormal loadings in **Q**. Given **Z** the $N \times A$ matrix of scores, defined by Zou et al.as

$$\mathbf{Z} = \mathbf{XP}, \tag{6}$$

a least squares approximation for **X** can be stated in the general form

$$\mathbf{X} = \mathbf{ZR}^{\mathrm{T}} + \mathbf{E}, \tag{7}$$

for **R** a $J \times A$ matrix satisfying

$$\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}, \tag{8}$$

where the superscript^{-1} indicates matrix inversion. Substituting $\mathbf{Z} = \mathbf{XP}$ into (8) we obtain

$$\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{XP}(\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XP})^{-1}. \tag{9}$$

At convergence, Eq. (5) holds for the found **P**, thus substituting (5) into (9) we get

$$\mathbf{R} = \mathbf{USV}^{\mathrm{T}}(\mathbf{P}^{\mathrm{T}}\mathbf{USV}^{\mathrm{T}})^{-1}. \tag{10}$$

Post-multiplying both members of (10) first by $\mathbf{P}^{\mathrm{T}}\mathbf{USV}^{\mathrm{T}}$ and then by $\mathbf{VS}^{-1}\mathbf{V}^{\mathrm{T}}$ we arrive at:

$$\mathbf{RP}^{\mathrm{T}}\mathbf{UV}^{\mathrm{T}} = \mathbf{UV}^{\mathrm{T}}. \tag{11}$$

At convergence, also Eq. (4) holds, thus substituting $\mathbf{Q} = \mathbf{UV}^{\mathrm{T}}$ into (11) and post-multiplying by $(\mathbf{P}^{\mathrm{T}}\mathbf{Q})^{-1}$ we arrive to an expression for **R** in terms of **P** and **Q**:

$$\mathbf{R} = \mathbf{Q}(\mathbf{P}^{\mathrm{T}}\mathbf{Q})^{-1}. \tag{12}$$

Finally, plugging (6) and (12) into (7) gives us the expression for the SPCA-Z model (at convergence) as function of **X**, the sparse weights **P** and orthonormal loadings in **Q** :

$$\mathbf{X} = \mathbf{XP}(\mathbf{Q}^{\mathrm{T}}\mathbf{P})^{-1}\mathbf{Q}^{\mathrm{T}} + \mathbf{E}. \tag{13}$$

3.2. Definition and interpretation of the scores

Eq. (13) can be broken down into equivalent matrix factorizations of **X**, with different interpretational properties, all of them of the form:

$$\mathbf{X} = \mathbf{A}_h\mathbf{B}_h + \mathbf{E}. \tag{14}$$

If we set $\mathbf{A}_1 = \mathbf{XP}$ and $\mathbf{B}_1 = (\mathbf{Q}^{\mathrm{T}}\mathbf{P})^{-1}\mathbf{Q}^{\mathrm{T}}$, we maintain the sparse weights and we get scores in $\mathbf{A}_1$ that are directly connected to them and so they are nicely interpretable. Unfortunately, components in $\mathbf{B}_1$ will likely be non orthogonal, and the interpretation of pairs or groups of components may be misleading.

If we set $\mathbf{A}_2 = \mathbf{XP}(\mathbf{Q}^{\mathrm{T}}\mathbf{P})^{-1}$ and $\mathbf{B}_2 = \mathbf{Q}^{\mathrm{T}}$, we give up on sparse weights as $\mathbf{P}(\mathbf{Q}^{\mathrm{T}}\mathbf{P})^{-1}$ is not sparse, and the scores $\mathbf{A}_2$ lose their direct interpretation from the sparse weights, but due to the orthogonality of $\mathbf{B}_2$ we can safely interpret pairs/groups of components. Furthermore, it turns out that $\mathbf{A}_2$ are the scores that best approximate the Euclidean distance between the observations in **X**. This will be shown in the remaining of this section.

For standardized variables, the Pearson's correlation coefficient and the Euclidean difference are functionally and inversely related.[2] Thus, finding the scores that best approximate the matrix of the correlation among observations $\mathbf{XX}^{\mathrm{T}}$ is equivalent to finding the scores that best approximate the Euclidean distance. This leads to the following minimization problem:

$$\underset{\mathbf{A}_h}{\operatorname{argmin}} \|\mathbf{XX}^{\mathrm{T}} - \mathbf{A}_h\mathbf{A}_h^{\mathrm{T}}\|_F^2. \tag{15}$$

In particular, the Euclidean distance for $\mathbf{A}_2$ holds

$$\|\mathbf{XX}^{\mathrm{T}} - \mathbf{XP}(\mathbf{Q}^{\mathrm{T}}\mathbf{P})^{-1}(\mathbf{P}^{\mathrm{T}}\mathbf{Q})^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\|_F^2, \tag{16}$$

and using the fact that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$ we arrive at

$$\begin{aligned}&\|\mathbf{XX}^{\mathrm{T}} - \mathbf{XP}(\mathbf{Q}^{\mathrm{T}}\mathbf{P})^{-1}\mathbf{I}(\mathbf{P}^{\mathrm{T}}\mathbf{Q})^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\|_F^2 = \\ &\|\mathbf{XX}^{\mathrm{T}} - \mathbf{XP}(\mathbf{Q}^{\mathrm{T}}\mathbf{P})^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}(\mathbf{P}^{\mathrm{T}}\mathbf{Q})^{-1}\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\|_F^2.\end{aligned} \tag{17}$$

Comparing the second equation in (17) with Eq. (7) we can re-write the former as

$$\|\mathbf{XX}^{\mathrm{T}} - \mathbf{XPR}^{\mathrm{T}}\mathbf{RP}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\|_F^2. \tag{18}$$

Formula (18) corresponds to the least squares solution for the approximation of $\mathbf{XX}^{\mathrm{T}}$ given **P**, and thus we can conclude that the scores $\mathbf{A}_2$ provide the best approximation in the low dimensional space of the distance between observations in the original data space.

The definition of the scores as $\mathbf{A}_2 = \mathbf{XP}(\mathbf{Q}^{T}\mathbf{P})^{-1}$ explicitly brings back the relevance of the orthonormal loadings in **Q**: we will now show that this has interesting consequences for the interpretation of the sparse components. It is worth noting the connection of these scores (and of the SPCA-Z model itself in Eq. (13)) with the well-known Partial Least Squares (PLS) regression model [14,15], where the sparse weights in SPCA-Z take the role of the weights in PLS, and the orthonormal loadings in SPCA-Z take the role of the so-called PLS loadings.

3.3. The role and interpretation of the orthonormal loadings

Zou et al.use the orthonormal loadings in **Q** as a mere computational trick. Yet, we have previously shown that they are an inherent part of the true underlying SPCA-Z model, and that they can provide scores with some desirable properties. For this reason, we attempt a characterization of the properties of **Q**.

To assess **Q** at convergence, *i.e.,* when the sparse weights **P** have been found, we reconsider problem (3). Using the equivalence of the trace operator tr and the Frobenius norm, the minimization problem

[2] For two standardized variables x and y measured on N replicates it holds that $\mathrm{corr}(x,y) = 1 - \frac{d^2(x,y)}{2N}$, with $\mathrm{corr}(x,y)$ and $d^2(x,y)$ are the Pearson's correlation and the Euclidean distance between x and y, respectively.

can be re-expressed in terms of the trace of the product of the matrices appearing in Eq. (3)

$$\underset{\mathbf{Q}}{\operatorname{argmin}} \left\{ \operatorname{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) + \operatorname{tr}(\mathbf{Q}\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}\mathbf{Q}^{\mathrm{T}}) - 2\operatorname{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}\mathbf{Q}^{\mathrm{T}}) \right\}. \tag{19}$$

Using the fact that the trace is invariant under cyclic permutation and that at convergence $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$, equation (19) can be expressed as:

$$\underset{\mathbf{Q}}{\operatorname{argmin}} \left\{ \operatorname{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) + \operatorname{tr}(\mathbf{P}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}) - 2\operatorname{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P}) \right\}. \tag{20}$$

Since the first two terms of Eq. (20) do not depend on $\mathbf{Q}$, minimizing it amounts to maximizing $\operatorname{tr}(\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{P})$. This can be interpreted as finding those orthonormal vectors in $\mathbf{Q}$ that maximize the covariance between the scores $\mathbf{Z} = \mathbf{X}\mathbf{P}$ and the auxiliary scores given by $\mathbf{X}\mathbf{Q}$.

When the minimization problem is rephrased in this way, we believe that the actual role of $\mathbf{Q}$ becomes apparent: to provide auxiliary scores mostly correlated with $\mathbf{Z}$ and the role of $\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}$ is to capture all the (possible) variance related to it. Hence, the orthonormal loadings in $\mathbf{Q}$ should not be considered just as a computational convenience, but rather as a fundamental part of the sparse model.

3.3.1. Interpreting sparse solutions in terms of representatives and associates

The final goal of sPCA modeling is to arrive at simpler and more interpretable models, where interpretability is usually gauged in terms of the (limited) number of non-zero weights. The non-zero weights for a given sparse principal component identify the variables that *represent* the structure of a sparse component, and for this reason we term them *representatives*.

On the other hand, $\mathbf{Q}$ represents a subspace in the rowspace of the data, and maintains the SPCA-*Z* model in that subspace. Therefore, orthonormal loadings complement $\mathbf{P}$ for the competing objective of maximizing variance in $\mathbf{X}$. In practice, this means that loadings $\mathbf{Q}$ are responsible for modeling the variance in the data that is correlated to $\mathbf{P}$. To provide an intuitive interpretation of this fact, we use the term *associates* to designate the variables with zero weight but a loading significantly different to zero in a given component. That is, variables that are correlated to the component score. While weights in $\mathbf{P}$ only contain *representatives*, the corresponding loadings in $\mathbf{Q}$ contain both *representatives* and *associates*.

To understand the connection between *representatives* and *associates*, the interpretation of the sparse factorization must be carried over simultaneously at both the sparse weights $\mathbf{P}$ and orthonormal loadings in $\mathbf{Q}$, because considering only the sparse weights can result in the loss of important information. Note that, for instance, as we force a component towards sparseness, some of the *representatives* are discarded from $\mathbf{P}$ but they may still being captured in the corresponding $\mathbf{Q}$. This is also related to another feature of the SPCA-*Z* approach: the potential non-uniqueness of the solution. Park et al. [8] have shown that when the number of variables is larger than the number of observations, different sets of variables can results in the equivalent sparse solutions (see Eqs. (8) and (9) in Appendix of that paper.). In those equivalent solutions, some of the *representatives* are exchanged with some *associates* giving the same exact data approximation. Thus, care should be taken in considering the *representatives* more relevant than the *associates* in interpretation.

We advocate for a broader and more comprehensive interpretation of the SPCA-*Z* solution that includes the *associates*. When interpreting a sparse model it should be possible to unequivocally identify the subset of variables connected to a given component. This problem is particularly important when sPCA is used for data exploration with the goal of finding groups of variables associated with the data patterns observed, like in the case of biomarker discovery. This is consistent with other sparsity factorizations based on the identification of groups of correlated variables [16].

3.4. Model selection using explained variance

An inherent problem of model fitting procedures where regularization parameters are involved is the choice of the *optimal* values for those parameters given the data to be analyzed, where optimality has to be defined in terms of some desired model properties. In the framework of sPCA this translates to finding the optimal trade-off between sparsity and variance of the solutions, including the determination of the number of components.

In the original publication, Zou et al. [6] suggest an interesting approach to set the level of sparsity when analyzing an experimental data set. They consider the trade-off between the sparsity in a component (*i.e.,* number of non-zero weights in $\mathbf{P}$) and the proportion of variance captured by that component. Taking as reference the explained variance by the components of a normal PCA model, they require the weights $\mathbf{P}$ to be as sparse as possible without reducing too much the variance with respect to the reference (this is shown in Figure 2 of [6]).

The rationale behind this criterion is that if a sparse model provides a reasonable factorization of the data (which happens if the data is inherently sparse with respect to an sPCA decomposition), the proportion of explained variance by the model should be large, where large can mean comparable to that captured by a standard PCA model. Then, if the data does not conform to an sPCA model, the variance captured by the model is expected to be low.

Zou et al.calculate the total explained variance by a sparse model with A components only considering the sparse weights $\mathbf{P}$, using the QR decomposition of $\mathbf{Z} = \mathbf{X}\mathbf{P}$. This calculation underestimates the true amount of explained variance by the sparse model if only $\mathbf{P}$ is considered [4]. Yet, if $\mathbf{Q}$ is also considered, as follows from the true underlying model in SPCA-*Z* , the variance is even larger. The actual model for the SPCA-*Z* is given by Eq. (13), for which the total explained variance follows

$$Tot V_{pq_A} = \operatorname{tr}(\mathbf{B}_2\mathbf{A}_2^{\mathrm{T}}\mathbf{A}_2\mathbf{B}_2^{\mathrm{T}}). \tag{21}$$

where the subscript pq makes explicit that the model variance makes use of both sparse weights and orthonormal loadings, and where we intentionally used the factorization following Eq. (14) that retains orthonormal loadings to allow for the computation of component-wise variance.

The underestimation of the explained variance has a profound impact on model selection. As a result, the desired trade-off between sparsity and explained variance is biased towards the latter, leading to suboptimal solutions with a number of non-zero weights much larger than what can be actually obtained without decreasing significantly the amount of explained variance. We will show this in Section 4, contextually to the analysis of an experimental data set.

3.5. Model selection in multi-component sPCA models using the razor plot

The selection of metaparameters in sPCA multi-component models is a real computational and practical challenge [16], since reducing the sparsity of the components (allowing for more non-zero weights) or adding more components are viable ways to increase the explained variance, but they affect the interpretability of a model differently. Besides, each component of a model may be fit with a different level of sparsity.

As the goal of sPCA is often exploratory, and given that the explained variance is an established criterion for the assessment of the quality of sPCA [17], it seems reasonable to follow the criterion suggested by Zou et al.: finding the sparsest model with explained variance close to an unconstrained PCA model. For this purpose, we make use of the razor plot [18].

The razor plot is a (multidimensional) extension of the scree-plot, where the number of parameters to be estimated is summarized as the number of free parameters [19]. In the case of sPCA, both the number

of components A and the number of non-zero weights per component nz_a need to be optimized. We define the optimization criterion f as

$$f = \sum_a nz_a - A. \tag{22}$$

where nz_a is the number of non-zero weights in $\mathbf{p}_a$. f takes values between 0 (all component have a single non-zero weight) and $(J-1)\cdot A$.

To compute a razor plot, sPCA models are fit for different values of the meta-parameters[3] and the explained variance is calculated for each model. This is a computationally intensive operation if we want to consider all possible combinations in the number of non-zero weights for the components. The computational burden is simplified if, without loss of generality, we restrict the number of non-zero weights in the a_2th component to be less or equal to the number of non-zero weights in the a_1th component for $a_1 < a_2$.

Once all model variants are computed, they are grouped by the value of f. The best model in terms of explained variance is chosen from the group, and the rest are discarded. The razor plot shows the explained variance of the set of best models as a function of f. An alternative plot that we introduce in this paper selects the best model for each combination of f and A, and visualizes the corresponding explained variance as a surface. In any of the two visualizations, our selection criterion is to pick the sparsest model (lowest f) for which the explained variance is close enough to the reference, where the reference is the explained variance of the corresponding PCA model (or any other target reference that can be deemed appropriate; e.g., the total variance in the data) and 'close enough' can be determined by a suitable threshold.

The rationale in the definition of f in Eq. (22), in which A is subtracted, is motivated by the fact that interpretability is favored by including more components for a fixed total number of non-zero weights. To give an example: a model with six components with one non-zero weight each is easier to interpret that a model with two components with three non-zero weights each. This observation agrees with sparse modeling approaches that try to avoid uncorrelated variables with high loadings in the same component [16,18].

Building a razor plot can be time consuming if the number of variables is very large and the parameter space to search is extensive. We noticed that if a data set complies with an sPCA structure, that is, if a relevant part of its variance can be captured by a set of sparse components, we do not need to explore the whole range of values of f. Rather, we can make a search that consistently assesses all models in increasing order of f and stops when the difference between the explained variance and the reference variance is smaller than a given threshold, say 5%. We have experimentally seen that if an sPCA model is suitable for a data set, we can compute this type of 'truncated' razor plot extremely quickly, while providing essentially the same information shown in complete razor plots. If the computation does not stop after a reasonable amount of time, this indicates that the sparse model is probably not a good modeling choice.

4. Putting sparse principal component analysis at work

Equipped with the theoretical results derived in the previous sections, we present a re-analysis of the so-called "pitprops" data set on which the functioning of SPCA-*Z* was first illustrated [6]. The goal is to show how the use of sparse weights **P** and the orthonormal loadings in **Q** as *representatives* and *associates* leads to better (more interpretable) sparse models. Subsequently, we revisit some of the results of Guerra-Urzola et al. [10], which showed limited performance of the sPCA solution by Zou et al.under certain conditions. All results are easily reproducible using the software described in Section 6, which includes Matlab routines to facilitate the application of sPCA to other data sets.

[3] Including λ_2 in Eq. (1), which does not affect f in Eq. (22).

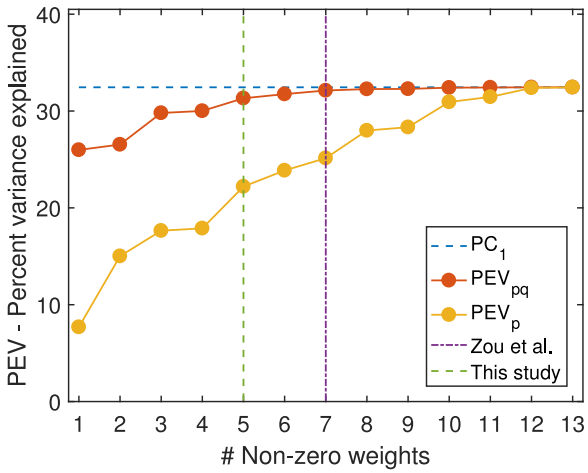


Fig. 1. Model calibration using explained variance. Proportion of explained variance (PEV) as function of model sparsity (*i.e.,* number of non-zero weights in **P**) for the first component of the sPCA model fitted to the pitprops data. Variance is calculated with approach of Zou et al.using only the sparse weights **P**, PEV_p, and Eq. (21) using both the sparse weights **P** and orthonormal loadings in **Q**, PEV_{pq}.

4.1. Analysis of the pitprops data

The data set contains 180 measurements of 13 functional variables measured on pitprops (lengths of lumber used to prop up the roofs of coal mines tunnels). First proposed by Jeffers [20], this data set has been widely used in the sPCA literature as a benchmark (see, for instance, [6,11,21–24]).

The sparse weights **P** and the orthonormal loadings in **Q** obtained by fitting a six-dimensional ($A = 6$) SPCA-*Z* model to the data are given in Table 1. We choose six components for consistency with Zou et al.and because most authors tend to agree about the relevance of the first six components for this data set (see [25], Chapter 8.7.1). We set the ridge penalty to 0 also for consistency with Zou et al.To the best of our knowledge, only the correlation matrix is known for this data set, so actual scores will not be discussed for data interpretation.

The explained variance by the first six sparse components, calculated both with the approach of Zou et al.and with Eq. (21), is given in Table 2, together with the explained variance by a standard PCA model as a reference.

The original approach by Zou et al.which only includes the sparse weights **P** underestimates the explained variance, in line with the discussion in the first paper of this series [4]. The explained variance by the full model is given by Eq. (21) and, as shown in Table 2, is larger than the explained variance using only the **P**: this happens because the orthonormal loadings in **Q** contribute to the variance calculation. The total variance is actually very close to the one captured by a PCA model with six components.

This has relevant implications for model calibration, as the use of the full model permits to explain a larger amount of variance than what was computed by Zou et al.with the same level of non-zero elements. Again, this means that we can probably arrive at a sparser, and potentially more interpretable, model than what Zou et al.originally determined.

4.1.1. Model calibration

The calibration of the level of sparsity (number of non-zero weights) is a fundamental step to obtain a meaningful sPCA model. For the model fit to the pitprops data, Zou et al.proposed seven non-zero weights for the first component. As described in Section 3.4 they proposed to explore the explained variance (per component) as a function of the number of non-zero weights.

Fig. 1 shows the explained variance by the first component, calculated with the approach of Zou et al.and with Eq. (21). We can observe

Table 1
Sparse weights **P** (left side) and orthonormal loadings in **Q** (right side) for a six-dimensional sPCA model fitted to the pitprop data represented by the correlation matrix among 13 variables measured on lumber lengths [20]. Empty cells indicate zero weights. The **P** table is the same as Table 3 in [6], where **P** are termed loadings.

Variable	Sparse weights **P**						Orthonormal loadings in **Q**					
	sPC1	sPC2	sPC3	sPC4	sPC5	sPC6	sPC1	sPC2	sPC3	sPC4	sPC5	sPC6
topdiam	−0.477						−0.460	0.110	0.016	−0.051	−0.075	0.166
length	−0.476						−0.475	0.057	0.022	−0.054	−0.078	0.188
moist		0.785					−0.032	0.709	−0.138	0.007	0.024	−0.053
testsg		0.619					0.029	0.676	0.139	0.023	0.039	−0.014
ovensg	0.177		0.641				0.212	−0.018	0.623	−0.008	0.118	0.123
ringtop			0.589				−0.043	0.066	0.597	−0.033	−0.132	−0.033
ringbut	−0.250		0.492				−0.243	−0.069	0.459	0.022	0.001	−0.113
bowmax	−0.344	−0.021					−0.349	−0.095	−0.061	−0.083	0.173	−0.122
bowdist	−0.416						−0.421	−0.021	−0.010	−0.084	−0.035	−0.022
whorls	−0.400						−0.395	−0.075	0.027	0.269	0.132	−0.151
clear				−1			0.003	−0.003	−0.003	−0.951	0.017	−0.026
knots		0.013			−1		0.004	−0.006	−0.013	0.022	−0.952	−0.033
diaknot			−0.016			1	−0.002	−0.010	−0.023	0.028	0.022	0.930

Table 2
Explained variance per component for a six-dimensional sPCA model fit to the pitprops data. PCA refers to explained variance by the successive components of a standard PCA model.

Percent of explained variance per component			
sPC	Zou et al.	Eq. (21)	PCA
1	28.0	28.1	32.5
2	14.0	15.5	18.3
3	13.3	15.6	14.4
4	7.4	8.6	8.5
5	6.8	8.6	7.0
6	6.2	8.8	6.3
Total	75.8	85.2	87.0

that, if the full model is taken into account to calculate variance, with a single non-zero (sparse) weight more variance is captured than what was estimated with the approach of Zou et al.for seven non-zero weights. The variance plot in Fig. 1 clearly shows the benefit of using the true model for calibration: even for a somehow conservative choice of the level of sparsity in this single component (five non-zero elements, two less than the choice of Zou et al.) it explains 31% of the variance, which is very close to the variance of the first principal component.

4.1.2. A closer look to the relationship between weights and loadings

In Section 3.3 we have discussed the role of the **Q** to maximize the correlation between the scores $\mathbf{Z} = \mathbf{XP}$ and the auxiliary scores given by **XQ**. Because of the way **P** and **Q** are constructed, with the weights **P** being sparse while **Q** is not, this correlation is dependent on the sparsity level. This is shown in Fig. 2 for the first component at different sparsity levels and where the correlation between $\mathbf{Xp}_1$ and $\mathbf{Xq}_1$ is shown. For a sparse model with just one non-zero element in $\mathbf{p}_1$ (Fig. 3(a)), the correlation between scores is low. This is related to how variance is captured in the model: the more sparsity, the less variance can be described by **P** and thus less correlation between $\mathbf{Xp}_1$ and $\mathbf{Xq}_1$. This situation of maximum sparsity corresponds to the leftmost value of PEV_{pq} in Fig. 1, with $PEV_{pq} = 26\%$. For a sparsity level of five non-zero weights in $\mathbf{p}_1$ there is a much stronger correlation (Fig. 3(b)) and larger explained variance ($PEV_{pq} = 31\%$, Fig. 1). For a non-sparse model, *i.e.,* a standard PCA model, the correlation between $\mathbf{Xp}_1$ and $\mathbf{Xq}_1$ (Fig. 3(c)) is perfect since in this case weights and loadings coincide ($\mathbf{P} = \mathbf{Q}$), and the explained variance is maximized.

4.1.3. Interpretation of a single sparse component: representatives and associates

Let us follow with the interpretation of a single sPCA component with five non-zero weights in terms of *representative* and *associates*. Fig. 3(a) and (b) show the sparse weights $\mathbf{p}_1$ and orthonormal loadings $\mathbf{q}_1$ for the first component, respectively. The non-zero weights in $\mathbf{p}_1$ correspond to the *representatives*: `topdiam`, `length`, `ringbut`, `bowdist` and `whorls`. The associated orthonormal loadings $\mathbf{q}_1$ have also high values in correspondence with these variables: in fact we have seen in Fig. 2(c) that there is a relatively high correlation between $\mathbf{q}_1$ and $\mathbf{p}_1$. However, we also note that other two variables `ringtop` and `bowmax` have high orthonormal loadings: these are the *associates*, which are related with the *representatives* but have not been selected within the sparse weights. In Fig. 3(c) we show a correlation map of the variance captured by the component, which allows to interpret how variables associate in the component: the correlation between `ringtop` and `bowmax` and all the *representatives* in $\mathbf{p}_1$ is apparent. This plot is useful to identify if all *representatives* and *associates* are correlated or they form subgroups; an indication that the component may be broken down into a larger number of sparser components.

Recall that Zou et al.selected seven non-zero weights in the first component, as shown in Table 1. All the representatives in our component and the associate `bowmax` are also present in first component of the Zou's. Yet, the latter also includes `ovensg`, and does not include our other associate `ringtop`. Inspection of the correlation map of the original data in Fig. 3(d) reveals that this variable is not correlated with any of the other variables in the component, while `ringtop` is indeed correlated to `ringbut`. Again, only interpreting the sparse weights is not optimal.

4.1.4. Multi-component sPCA model selection for the pitprops data using the Razor plot

The razor plot for the pitprops data set in shown in Fig. 4. Fig. 4(a) shows the razor plot against the value of f. The plot shows that for $f > 6$, the gain in explained variance is small ($PEV_{pq} = 85.3\%$), and it is already almost equal to the one explained by a standard PCA model with six components ($PEV = 86\%$, orange bar in the plot), with a difference smaller than 1%. More insights may be obtained using a razor plot as function of f and the number of components, as shown in Fig. 4(b). This plot is more informative since it shows the interplay between model dimensionality and sparsity: an explained variance of almost 80% can be obtained with a model with six components with just one non-zero weight per component. This result shows that the pitprops data is an excellent example to apply sPCA, which probably explains its popularity.

The razor plot can also be used as a visual diagnostic to assess whether a data set is sparse in an sPCA sense. If the razor plot suggests that a large value of f (*i.e.,* a large number of non-zero weights and/or a large number of components) is needed to explain an amount of variance similar to that explained by a standard PCA model, then we may conclude that the data set being analyzed is not sparse and, as such, sPCA should probably not be used.

Generating the data in Fig. 4, exploring models up to $f = 78$, took approximately four hours on a standard desktop computer without parallelization. For very large data sets the problem becomes rapidly intractable. Truncated razor plots like those shown in Fig. 5 can be

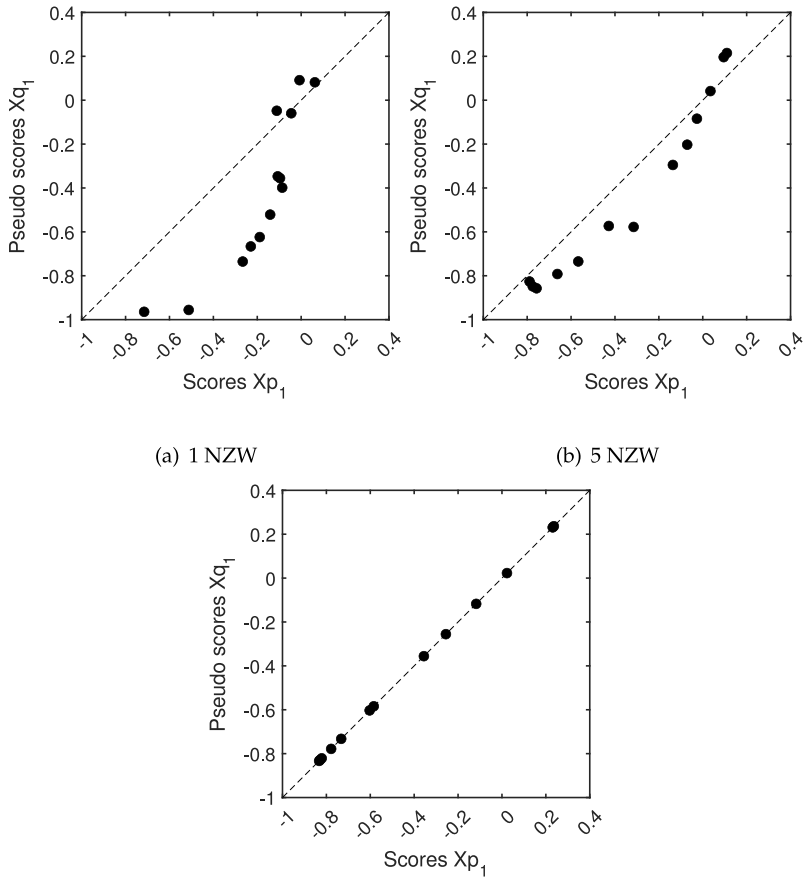


Fig. 2. Scores $\mathbf{Xp}_1$ vs. pseudo-scores $\mathbf{Xq}_1$ for the first component of the SPCA-*Z* model fitted to the pitprops data for different levels of sparsity: (a) one non-zero weight (NZW), (b) five non-zero weights, and (c) first PC (*i.e.,* thirteen non-zero weights).

obtained in less than one second, while providing essentially the same information shown in Fig. 4. The simpler model satisfying the threshold is a model with six components and two non-zero weights per component ($f = 6$), explaining 95% of the variance of a reference PCA model. Fig. 6 shows the sparse weights $\mathbf{p}_a$ and the orthonormal loadings $\mathbf{q}_a$ for the six components together with the correlation matrix of the factorization components.

It should be noted that this solution is different from the one shown in Table 1 and Fig. 3, because the model selection with the razor plot allows us to take into consideration the overall complexity of the model, rather than fitting one component at a time. Fig. 7 shows how combining the contributions of each sparse component (Fig. 7(a)) provides a good approximation of the sample correlation in the data $\mathbf{X}^T\mathbf{X}$ (Fig. 7(b)), indicating the appropriateness of the sparse model to fit this data set.

4.2. Revisiting the results of Guerra-Urzola et al. [10]

Guerra-Urzola et al. [10] performed a systematic comparison of several sparse PCA methods where they considered diverse data generation schemes that could lead to seemly similar sparse data. They report a poor performance of the sPCA version by Zou et al.under certain conditions: these results conflict with what we described in the first two papers of this series [4,5], where the performance and the characteristics of the Zou's algorithm were investigated using a data simulation strategy that is closer in philosophy to situations that are the most challenging sPCA. For this reason we believe it interesting to reproduce part of their experiments.

4.2.1. Simulation results

In the first part of their paper, Guerra-Urzola et al.present simulation results for three different types of conditions. We reproduce two of them here: (*i*) when the sparse structure of the data reflects the type of sparse structure addressed by sPCA (matching sparsity), thus generating data which should be easy to model, and (*ii*) when there is a mismatch between generated and estimated sparse structures (mismatching sparsity), resulting in a more challenging case. To compare the different models they define several performance indexes, including: the squared relative error (SRE) of the model parameters (SRE_LW for the sparse weights and SRE_S for the scores), which should be as low as possible; the percentage of explained variance (PEV), which should be close to, but not exceed, the variance accounted by the generated data (VAF); and the cosine similarity, which should be as close to one as possible (CS_LW for sparse weights and CS_S for scores). We reproduced the experiment in a configuration point of the simulation where the performance of sPCA was particularly poor: for two components, 500 observations, 1000 variables, 80% of sparsity and 80% of VAF (we provide the code that allows to reproduce any other configuration, see Section 6).

Table 3 shows the performance results of sPCA for two values of λ_2 in Eq. (1). For $\lambda_2 = 1$, the results are consistent with those by Guerra-Urzola et al.and thus very poor, but for $\lambda_2 = \infty$, the performance is competitive with the best sparse methods reported in their study. It turns out that, in our previous papers of this series, we always considered $\lambda_2 = \infty$ (i.e., soft-thresholding), since chemometrics data is often very wide with many more variables (columns) than observations (rows): this is the suggested λ_2 configuration by Zou et al.for this type

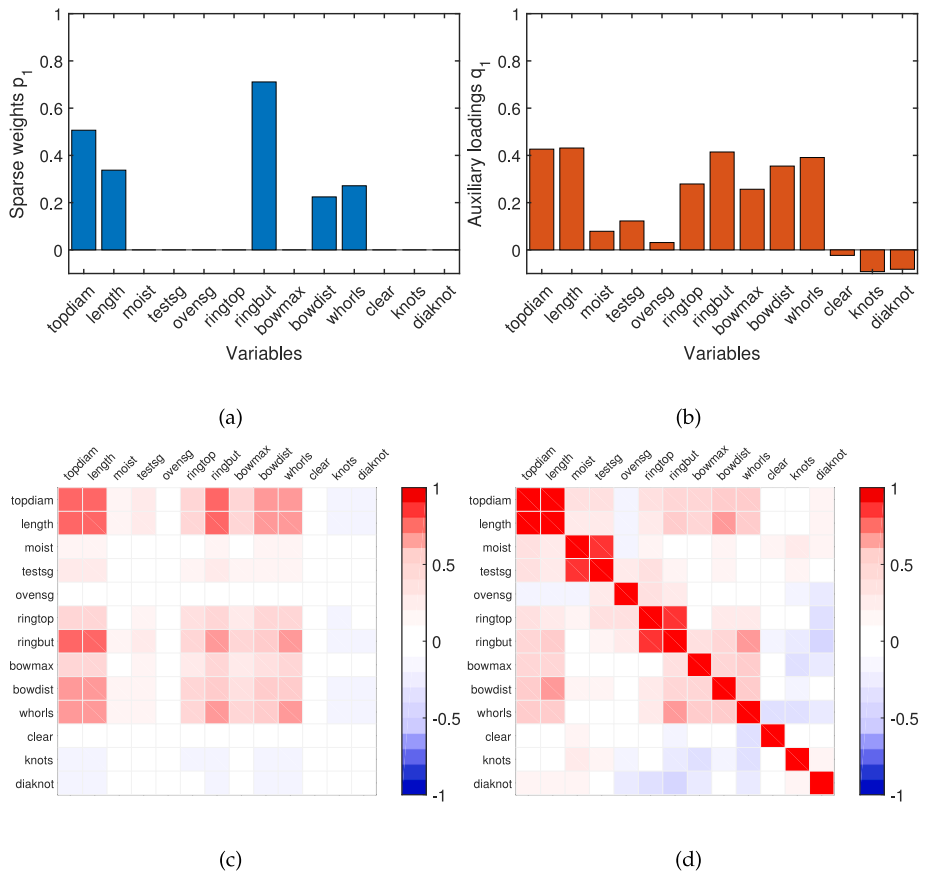


Fig. 3. Overview of the first component of the sPCA model for the pitprops data set: (a) sparse weights $\mathbf{p}_1$, (b) orthonormal loadings $\mathbf{q}_1$, (c) correlation matrix of the model factorization $\mathbf{r}_1\mathbf{p}_1^T\mathbf{X}^T\mathbf{X}\mathbf{p}_1\mathbf{r}_1^T$, and (d) data correlation matrix $\mathbf{X}^T\mathbf{X}$.

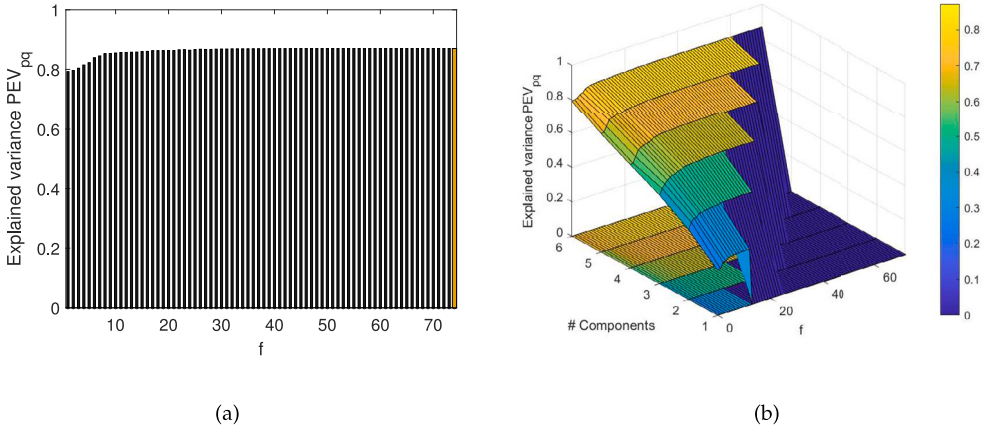


Fig. 4. Calibration of a multi-component sPCA model for the pitprops data: (a) razor plot with explained variance PEV_{pq} as function of the optimization criterion f (22); and (b) razor plot with explained variance PEV_{pq} as function of f and the number of sparse components. The explained variance by a standard PCA model with six components is given as a reference. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

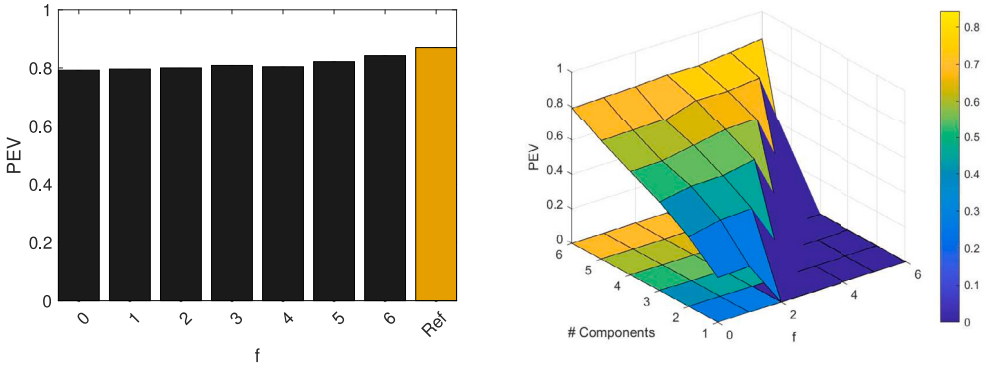
of data. This result shows that the choice of λ_2 can be relevant in certain examples, and that sPCA can still provide good performance for both matching and mismatching sparse data generation.

4.2.2. Multi-component sPCA model selection for the Big Five data using the Razor plot

Guerra-Urzola et al.also presented results for two real data sets. The first one is the Big Five personality dimensions data [26] publicly available from the R-package qgraph [27], which contains the scores of 500 individuals on five sets of 48 items associated to the Big Five personality traits (Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness). To assess the performance of sparse methods in this case study, Guerra-Urzola et al.considered an optimal result if "*most nonzero loadings* (or weights) *belong to one particular item set*" and all item sets were modeled.

We computed the razor plot for the following set of parameters:

- Numbers of non-zero weights considered per component, nz_a, from 1 to 240 (the total of variables) in steps of 24.
- Values of λ_2 considered: $[0, 1, 10, 100, 10000, \infty]$.



(a) (b)

Fig. 5. Calibration of a multi-component sPCA model for the pitprops data: (a) truncated razor plot with explained variance PEV_{pq} as function of the optimization criterion f (22); and b) truncated razor plot with explained variance PEV_{pq} as function of f and the number of sparse components. The explained variance by a standard PCA model with six components is given as a reference.

Table 3
Performance of the Zou's sPCA algorithm using the Guerra-Urzola et al. [10] simulation for two values of λ_2 in Eq. (1) and two conditions: matching sparsity and mismatching sparsity; Alg-I and Alg-II refer to the data generating procedures defined as Algorithm 1 and Algorithm 2 in [10]. Performance is expressed as squared relative error of the sparse weights (SRE_LW) and the scores (SRE_S), the percentage of explained variance (PEV) and the cosine similarity of sparse weights (CS_LW) and of scores (CS_S)).

λ_2	Matching sparsity (Alg-II)			Mismatching sparsity (Alg-I)		
	SRE_LW	SRE_S	PEV	CS_LW	CS_S	PEV
1	0.22	0.03	0.80	0.56	0.76	0.80
∞	0.17	0.03	0.80	1.00	1.00	0.80

- Threshold of 5% below the variance of a 5 PC model (24.75%).

Fig. 8(a) shows the truncated razor plot: the PEV increases fast up to $f = 120$ and reaches the threshold at $f = 216$. This plot was computed in less than five minutes. The selected model has five components, the first four with 49 nz_a (recall potential values start in one and increase in steps of 24), which is the closest situation to select one group of 48 item per component. The last component has 25 nz_a.

In this optimal configuration, all λ_2 values provide solutions with similar PEV, with $\lambda_2 = 10$ slightly outperforming the others, reason why this choice is selected. The optimal sparse configuration according to Guerra-Urzola et al.was for the sPCA-rSVD algorithm [23], with 0.73 of sparsity and 18% of PEV. In comparison, our sPCA configuration attains 0.82 of sparsity and 24% of PEV, with more captured variance and less non-zero parameters. Fig. 9 shows the sparse weights of this model, where each component is mainly associated to a single personality trait.

In Table 4, we compare the maximum ratio of nonzero loadings/weights in one particular item set in the sPCA-rSVD and sPCA models by Guerra-Urzola et al. [10] (results taken from Table 4 of their article) with our sPCA configuration. Following the criteria of Guerra-Urzola et al.our configuration outperforms traditional sPCA and the alleged optimal algorithm in terms of the percentage of association of one component to one personality trait (see the last row in the Table 4) and the PEV captured. Yet, the openness dimension is not properly recovered, which can be interpreted as a disadvantage or the need to add more components. As shown in Fig. 10A the reconstructed map $\mathbf{RP}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XPR}^{\mathrm{T}}$ computed from our model shows a characteristic diagonal block shape that justifies the correlation within traits, but also shows a tendency for negative correlation between Neuroticism and Extraversion, and other more moderate relations between traits, which are relevant for the full understanding of the data.

4.2.3. Multi-component sPCA model selection for the gene expression data using the Razor plot

The last example used by Guerra-Urzola et al.is a data set containing gene expression of 13 male autistic individuals (6 with autism caused by a fragile X mutation (FMR1-FM) and 7 with autism caused by a 15q11–q13 duplication (dup15q)) and 14 controls [28], available at NCBI GEO database with accession number GSE7329. The data sets contains 41 675 gene probes (Guerra-Urzola et al.report 43 893 probes, probably by mistakenly including some control probes).

In their analysis Guerra-Urzola et al.selected three components, for which PCA captures 32% of the variance. They selected the GPower method [11] as their reference sparse configuration for this data, with 0.97 of sparsity and 31% of PEV. We repeated the analysis with three components in PCA capturing 33% of the variance (this small discrepancy is probably due to the different number of variables considered).

We computed the razor plot for the following set of parameters:

- Numbers of non-zero weights considered per component, nz_a, taking into account the matrix rank (26) and the number of variables: $[1, 4, 7, 10, 13, 16, 19, 22, 25, 41675]$.
- Values of λ_2 considered: $[0, 1, 10, 100, 10000, \infty]$.
- Threshold of 3% below the variance of a 3 PC model (33%).

The computation of the truncated razor plot took 72 seconds, and all λ_2 values provided very similar results (we chose $\lambda_2 = 0$). The optimal output was 3 components with 4 non-zero weights each, for a total sparsity of 0.9999, capturing 32.6% of the variance: this model outperforms that obtained by Guerra-Urzola et al.in their optimal configuration. Note that Guerra-Urzola et al.did not run sPCA on these data due to the computational burden of the algorithm.

The component scores for the PCA and sPCA models are given in Fig. 11, showing a very similar separation between the three group of subjects. The ability to summarize data of sPCA may seem remarkable in this case study (like Guerra-Urzola et al.suggested for the GPower algorithm [11]), but in a closer inspection the auxiliary loadings (see Fig. 10B for an example) allows us to fully understand what is truly happening.

For each component, 4 non-zero weights are enough to model the data, but these are only *representatives* connected to a much higher number of other *associates*, which can (and some of them quite probably will) be more relevant to understand the underlying disease biology than the selected *representatives*. Clearly, this is the consequence of having a very large number of potential biomarkers measured on a very limited number of samples (41 675 versus 27, in this example).

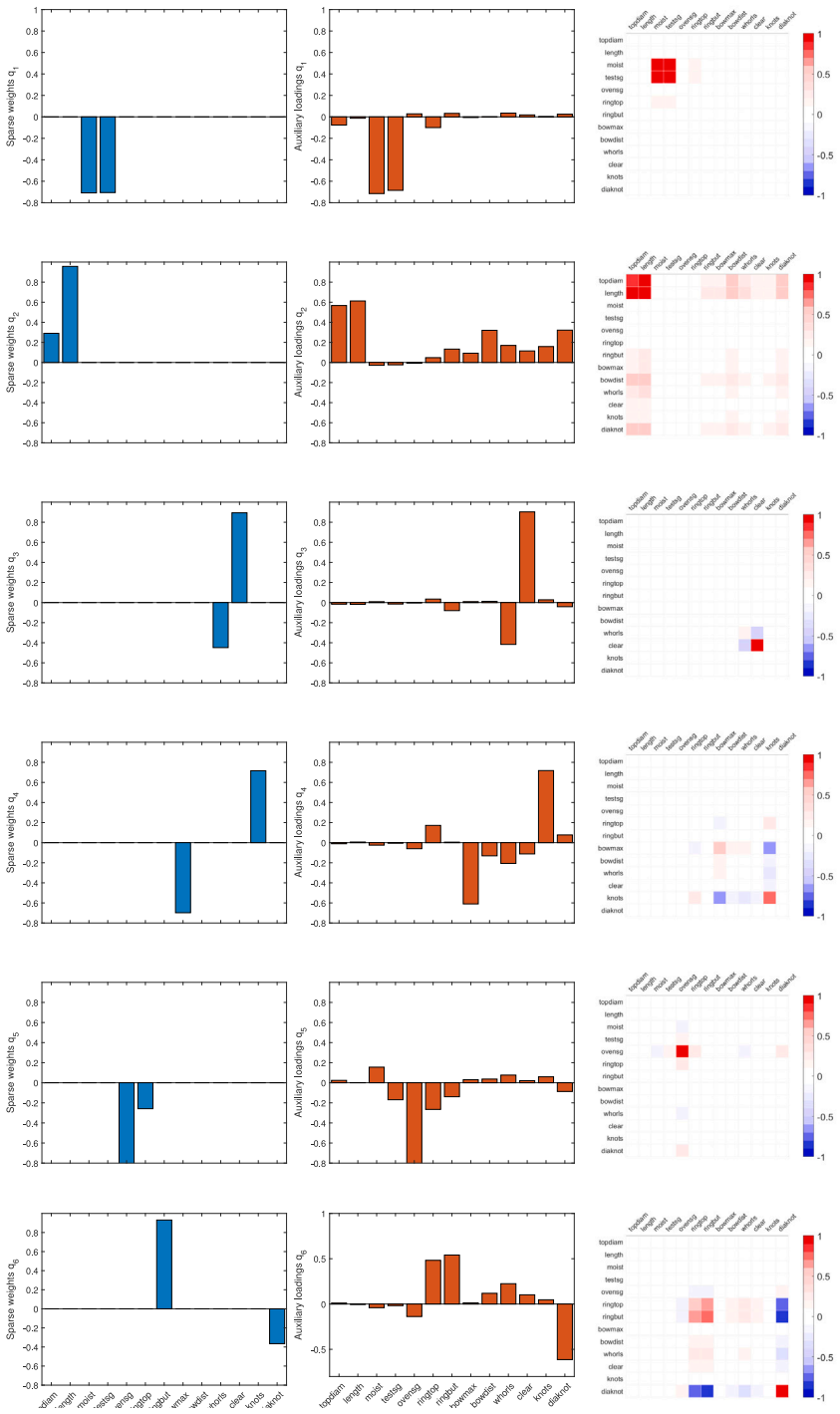
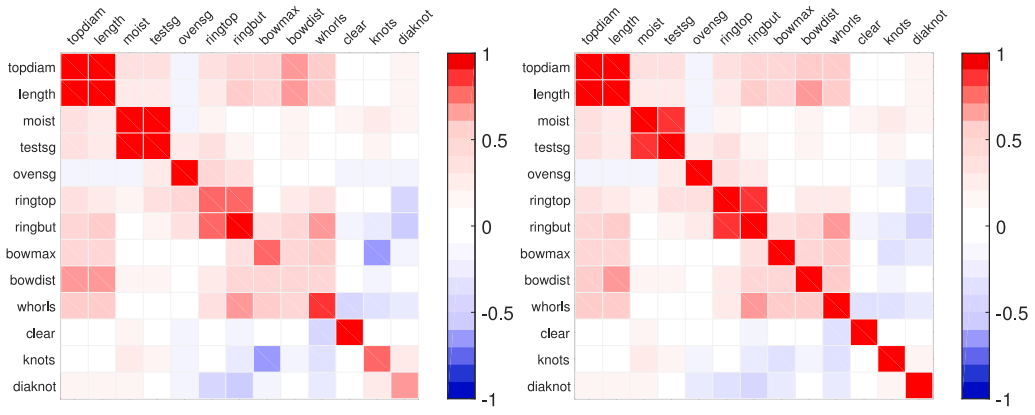


Fig. 6. Multi-component sPCA model for the pitprops data sets. Sparse weights $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_6$ and orthonormal loadings $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_6$ together with the corresponding visual map of correlation matrix of the model factorization $\mathbf{r}_j \mathbf{p}_j^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{p}_j \mathbf{r}_j^{\mathrm{T}}$ for each component. This model has been selected using the truncated Razor plot shown in Fig. 5.

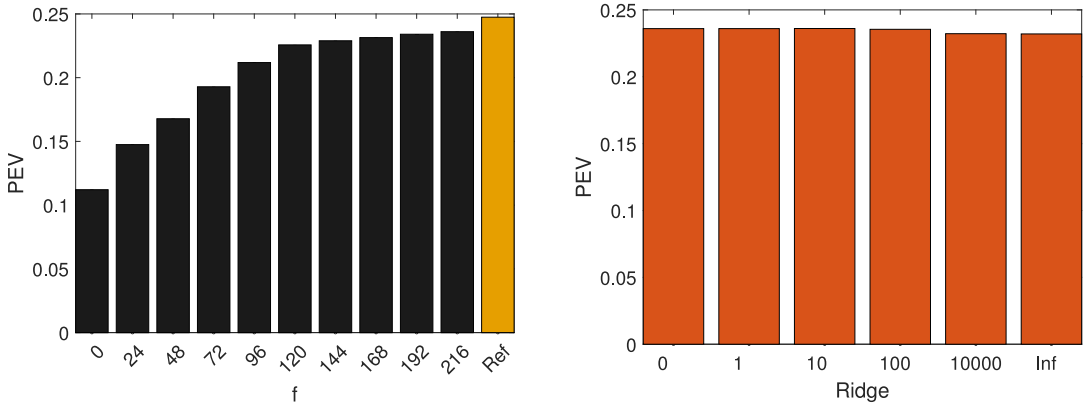
5. Conclusion

This series of three papers "All sparse PCA models are wrong, but some are useful" presents a combination of theoretical derivations, numerical simulations and real data analyses with the aim of improving our understanding on how to make the most of the interpretational capabilities of sparse PCA. The first paper showed that sPCA implementations underperform in situations where they were expected to provide a perfect solution. The second paper showed why deflation-based sPCA implementations are the most affected by model inaccuracies. This result made us select alternative implementations for the study in this third paper, and in particular we focused in the most popular implementation of sPCA, the one by Zou et al.. Here, we reviewed this version of sPCA, and provided new derivations on the model equations, the computation and interpretation of scores, the interpretation of weights and loadings and the selection of metaparameters. We believe that with these additions, sPCA becomes a more mature and well-understood method that can more safely be applied in critical analyses.



(a) (b)

Fig. 7. Correlation map of the variables of the pitprops given by (a) the reconstruction ($\mathbf{RP}^T\mathbf{X}^T\mathbf{XPR}^T$) with the six-dimensional model shown in Fig. 6 and (b) the data $\mathbf{X}^T\mathbf{X}$.



(a) (b)

Fig. 8. Truncated razor plot (a) and performance (PEV) of the group of optimal model variants with different λ_2 value (b) for the sparse PCA model of the Big Five personality data [26] (48 items measured on 500 individuals).

Table 4
Sparse PCA model of the Big Five personality data. Each column represents the number of items in each loading/weight that have a nonzero value in each trait. The components were ordered following Guerra-Urzola et al. [10], Table 4.

	SPCArSVD [10]					sPCA [10]					Our sPCA				
	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$
Openness	0	9	1	4	**41**	0	17	4	13	**25**	4	5	2	5	0
Conscientiousness	9	3	11	**43**	2	15	0	**26**	**24**	8	9	11	**22**	11	2
Extraversion	17	19	21	6	9	15	10	15	6	16	1	0	5	0	**23**
Agreeableness	4	**29**	**23**	2	5	6	**27**	13	10	3	**34**	5	5	5	0
Neuroticism	**34**	4	8	9	7	**28**	10	6	11	12	1	**28**	15	**28**	0
Total nonzero	64	64	64	64	64	64	64	64	64	64	49	49	49	49	25
% in one item set	**53**	**45**	**36**	**67**	**64**	**44**	**42**	**41**	**38**	**39**	**69**	**57**	**45**	**57**	**92**

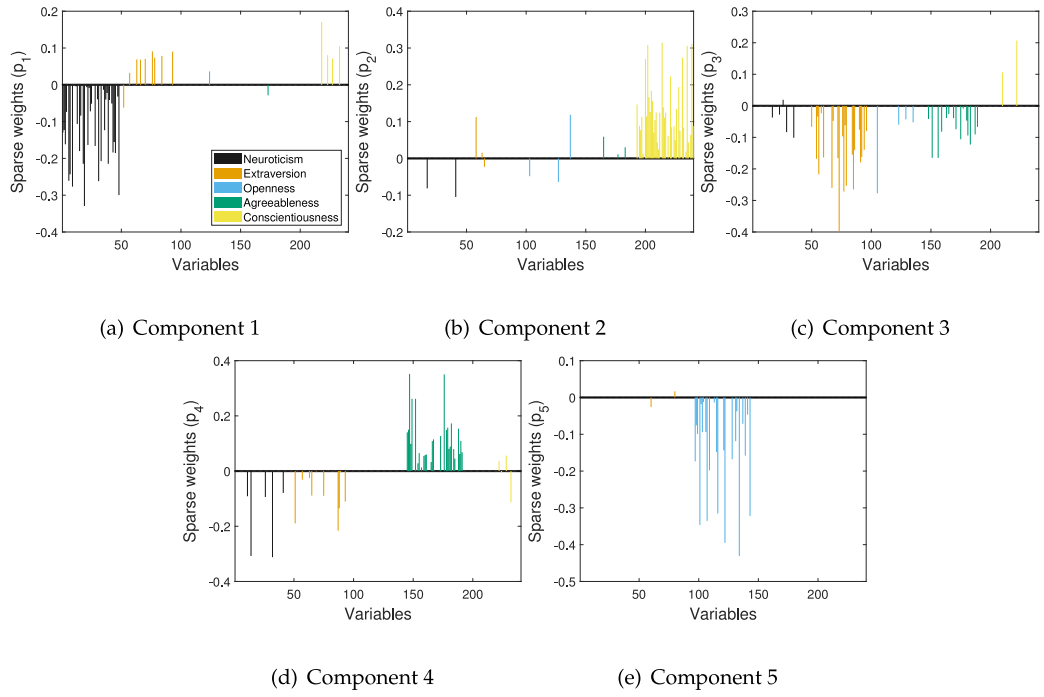


Fig. 9. Sparse weights for the 5 components of the sPCA model fitted to the Big Five personality data.

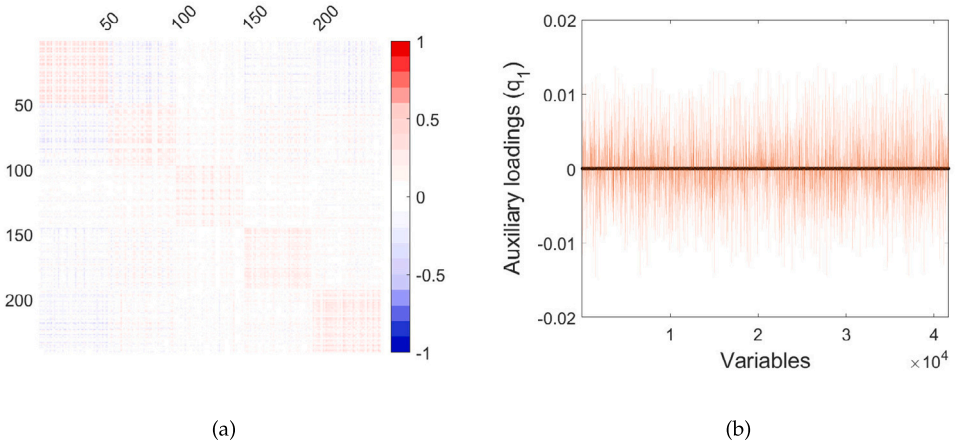


Fig. 10. (a) Reconstruction ($\mathbf{RP}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{XPR}^{\mathrm{T}}$) of the overall correlations patterns among the 240 original variables of the Five personality dimensions data given by the five-dimensional SPCA model. The order of traits is Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. (b) Auxiliary loadings for component 1 in sPCA Gene Expression Data. Non-zero sparse weights are NM_020156, NM_012247, NM_001008756, NM_000690.

6. Software

The results can be reproduced with the scripts at https://github.com/josecamachop/SparsePCAIII. The code makes use of the MEDA Toolbox v1.8 at https://github.com/codaslab/MEDA-Toolbox and the SPASM Toolbox at https://www2.imm.dtu.dk/projects/spasm. The MEDA Toolbox v1.8 also includes the implementation of sPCA based on soft-thresholding (routine spca) and the Razor plot (routine razorPlot) without dependence on any other software package. The script https://github.com/josecamachop/SparsePCAIII/pitpropsMEDA.m illustrates its use.

CRediT authorship contribution statement

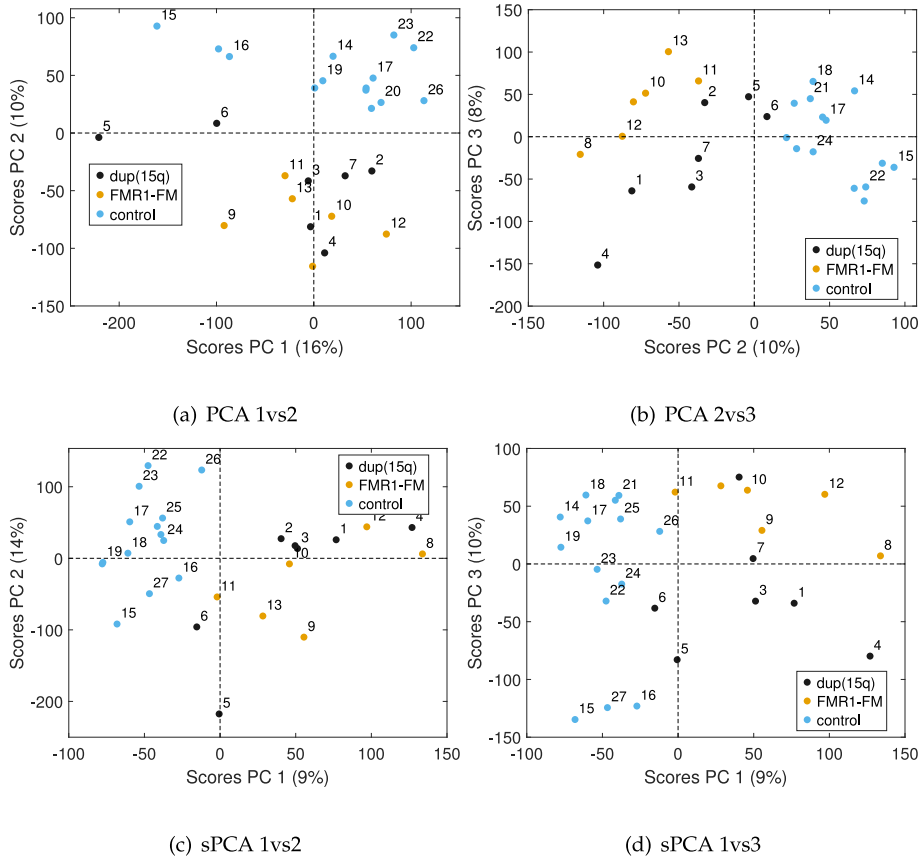
J. Camacho: Writing – original draft, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **A.K. Smilde:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **E. Saccenti:** Writing – original draft, Visualization, Conceptualization. **J.A. Westerhuis:** Writing – review & editing. **R. Bro:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by grant no. PID2023-1523010B-IOO (MuSTARD), funded by the Agencia Estatal de Investigación in Spain, call no. MICIU/AEI/10.13039/501100011033, and by the European Regional Development Fund, and the "Plan Propio de la Universidad



(a) PCA 1vs2 (b) PCA 2vs3

(c) sPCA 1vs2 (d) sPCA 1vs3

Fig. 11. Scatter plot of PCA scores (a)–(b) and of sPCA scores (c)–(d) for the in the Gene Expression Data. FMR1-FM indicates subjects with autism caused by a fragile X mutation; dup15q indicates subjects with autism caused by a 15q11–q13 duplication.

de Granada". Funding for open access charge: Universidad de Granada / CBUA

Appendix. Other sPCA algorithms related to SPCA-Z

In the previous papers of the series, we found three sPCA implementation generally superior to the rest, which were deflation-based. The SPCA-Z algorithm is one of them. The second is a sequential variant of SPCA-Z, in which the loadings are computed one at a time, published as part of the SPASM Toolbox [7]. This approach combines deflation with the use of orthonormal loadings, speeding up computation at the expense of a reduction in explained variance. This reduction represents a loss of modeling accuracy. The third is a variant of the popular Penalized Matrix Decomposition (PMD) by Witten et al. [2]. The variant is an alternative to projection deflation by computing orthogonalized scores. Yet, authors claim that it is not clear whether orthogonality is a desirable property.

We recently showed that orthogonal scores actually improve the estimation of sPCA by avoiding severe departures of the loadings from the data rowspace [5]. While PMD with orthogonal scores is somehow similar to SPCA-Z, in the later we have the combination of sparse weights and orthonormal loadings in which we rely for interpretation. For these reasons, and given that deflation-based approaches are generally inferior to these three sPCA variants [5], we will focus the paper on SPCA-Z.

Data availability

A link to the code and data is provided in the paper.

References

[1] Lester Mackey, Deflation methods for sparse PCA, Nips (2008) 1–8.
[2] Daniela M. Witten, Robert Tibshirani, Trevor Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[3] Trevor Hastie, Robert Tibshirani, Martin Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, Chapman & Hall/CRC, 2015.
[4] José Camacho, Age K. Smilde, Edoardo Saccenti, Johan 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.
[5] José Camacho, Age K. Smilde, Edoardo Saccenti, Johan A. Westerhuis, Rasmus Bro, All sparse PCA models are wrong, but some are useful. Part II: Limitations and problems of deflation, Chemometr. Intell. Lab. Syst. 208 (2021) 104212.
[6] Hui Zou, Trevor Hastie, Robert Tibshirani, Sparse principal component analysis, J. Comput. Graph. Statist. 15 (2) (2006) 265–286.
[7] Karl Sjöstrand, Line Clemmensen, Rasmus Larsen, Gudmundur Einarsson, Bjarne Ersbøll, SpaSM: A MATLAB toolbox for sparse statistical modeling, J. Stat. Softw. Artic. 84 (10) (2018) 1–37.
[8] S. Park, E. Ceulemans, K. Van Deun, A critical assessment of sparse PCA (research): why (one should acknowledge that) weights are not loadings, Behav. Res. Methods 56 (3) (2024) 1413–1432.
[9] Edoardo Saccenti, Johan A. Westerhuis, Age K. Smilde, Mariët J. van der Werf, Jos A. Hageman, Margriet M.W.B. Hendriks, Simplivariate models: uncovering the underlying biology in functional genomics data, PloS One 6 (6) (2011) e20747.
[10] Rosember Guerra-Urzola, Katrijn Van Deun, Juan C. Vera, Klaas Sijtsma, A guide for sparse pca: Model comparison and applications, Psychometrika 86 (4) (2021) 893–919.
[11] Michel Journée, Yurii Nesterov, Peter Richtárik, Rodolphe Sepulchre, Generalized power method for sparse principal component analysis, J. Mach. Learn. Res. 11 (2) (2010).
[12] H. Zou, T. Hastie, Regularization and variable selection via the elastic-net, J. R. Stat. Soc. 67 (2005) 301–320.
[13] Hui Zou, Lingzhou Xue, A selective overview of sparse principal component analysis, Proc. IEEE 106 (8) (2018) 1311–1320.

[14] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[15] H. Wold, E. Lyttkens, Nonlinear iterative partial least squares (NIPALS) estimation procedures, in: Bull. Intern. Statist. Inst. Proc., 37th Session, London, 1969, pp. 1–15.
[16] José Camacho, Rafael A. Rodríguez-Gómez, Edoardo Saccenti, Group-wise principal component analysis for exploratory data analysis, J. Comput. Graph. Statist. 26 (3) (2017) 501–512.
[17] I.T. Jolliffe, N.T. Trendafilov, M. Uddin, A modified principal component technique based on the LASSO, J. Comput. Graph. Statist. (2003).
[18] Jose Camacho, Evrim Acar, Morten A. Rasmussen, Rasmus Bro, Cross-product penalized component analysis (X-CAN), Chemometr. Intell. Lab. Syst. 203 (2020) 104038.
[19] Eva Ceulemans, Henk A.L. Kiers, Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method, Br. J. Math. Stat. Psychol. 59 (1) (2006) 133–150.
[20] John N.R. Jeffers, Two case studies in the application of principal component analysis, J. R. Stat. Soc. Ser. C. Appl. Stat. 16 (3) (1967) 225–236.
[21] Ian T. Jolliffe, Nickolay T. Trendafilov, Mudassir Uddin, A modified principal component technique based on the LASSO, J. Comput. Graph. Statist. 12 (3) (2003) 531–547.
[22] Baback Moghaddam, Yair Weiss, Shai Avidan, Spectral bounds for sparse PCA: Exact and greedy algorithms, Adv. Neural Inf. Process. Syst. 18 (2005).
[23] Haipeng Shen, Jianhua Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivariate Anal. 99 (6) (2008) 1015–1034.
[24] Yang Wang, Qiang Wu, Sparse PCA by iterative elimination algorithm, Adv. Comput. Math. 36 (2012) 137–151.
[25] I.T. Jolliffe, Principal Component Analysis, Springer Verlag Inc., EEUU, 2002.
[26] Paul T. Costa, Robert R. McCrae, A five-factor theory of personality, Handb. Pers.: Theory Res. 2 (01) (1999) 1999.
[27] Sacha Epskamp, Angélique O.J. Cramer, Lourens J. Waldorp, Verena D. Schmittmann, Denny Borsboom, Qgraph: Network visualizations of relationships in psychometric data, J. Stat. Softw. 48 (2012) 1–18.
[28] Yuhei Nishimura, Christa L. Martin, Araceli Vazquez-Lopez, Sarah J. Spence, Ana Isabel Alvarez-Retuerto, Marian Sigman, Corinna Steindler, Sandra Pellegrini, N. Carolyn Schanen, Stephen T. Warren, et al., Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways, Hum. Mol. Gen. 16 (14) (2007) 1682–1698.