PSYCHOMETRIKA
https://doi.org/10.1007/s11336-018-9650-9

SEMI-SPARSE PCA

LARS ELDÉN

LINKÖPING UNIVERSITY

NICKOLAY TRENDAFILOV

THE OPEN UNIVERSITY

It is well known that the classical exploratory factor analysis (EFA) of data with more observations than variables has several types of indeterminacy. We study the factor indeterminacy and show some new aspects of this problem by considering EFA as a specific data matrix decomposition. We adopt a new approach to the EFA estimation and achieve a new characterization of the factor indeterminacy problem. A new alternative model is proposed, which gives determinate factors and can be seen as a semi-sparse principal component analysis (PCA). An alternating algorithm is developed, where in each step a Procrustes problem is solved. It is demonstrated that the new model/algorithm can act as a specific sparse PCA and as a low-rank-plus-sparse matrix decomposition. Numerical examples with several large data sets illustrate the versatility of the new model, and the performance and behaviour of its algorithmic implementation.

Key words: alternative factor analysis, matrix decompositions, least squares, Stiefel manifold, sparse PCA, robust PCA.

1. Introduction

Let $z \in \mathbb{R}^{p\times 1}$ be a random vector of standardized observable variables. The classical EFA (exploratory factor analysis) model states that z can be written in the form (Mulaik, 2010, p. 136):

$$z = \Lambda f + \Psi u, \tag{1}$$

where $f \in \mathbb{R}^{m\times 1}$ is a random vector of m ($m \ll p$) common factors, $\Lambda \in \mathbb{R}^{p\times m}$ is a matrix of fixed factor loadings, $u \in \mathbb{R}^{p\times 1}$ is a random vector of unique factors and Ψ is a $p \times p$ diagonal matrix of fixed coefficients called uniqueness.

Thus, EFA tries to express p random variables by fewer m new variables and additional adjustment for each variable, i.e. involving another p unique variables (factors). EFA can be interpreted as an attempt to improve principal component analysis (PCA) by doing a similar low-rank approximation and polishing it by a slight adjustment of each fitting variable. However, that means that EFA expresses p observable variables by new *unknown* $m + p$ variables. Another baffling feature of the EFA model (1) is that it includes both random and nonrandom unknowns. One would expect some hybrid solution obtained by parametric and nonparametric methods.

Instead, the classical EFA takes the following approach. Let $\mathcal{E}$ denote the expectation operator. EFA assumes that $\mathcal{E}(f) = O_{m\times 1}$, $\mathcal{E}(u) = O_{p\times 1}$, $\mathcal{E}(uu^\top) = I_p$ and $\mathcal{E}(fu^\top) = O_{m\times p}$, where I_p is an identity matrix of order p and $O_{p\times m}$ is a $p\times m$ matrix of zeros. Furthermore, let $\mathcal{E}(zz^\top) = \Sigma$

Correspondence should be made to Nickolay Trendafilov, School of Mathematics and Statistics, The Open University, Milton Keynes, UK. Email: nickolay.trendafilov@open.ac.uk

Published online: 27 November 2018

be the population correlation matrix and $\mathcal{E}(ff^\top) = I_m$, i.e. assuming uncorrelated (orthogonal) common factors. Then, the population EFA m-model (1) and the assumptions made imply that

$$\Sigma = \Lambda\Lambda^\top + \Psi^2, \tag{2}$$

and the unknown random variables f and u luckily vanished.

In reality, we do not know the population correlation matrix Σ. We have either a $p \times p$ sample correlation matrix or a $n \times p$ data matrix Z collecting n independent observations on $p(< n)$ standardized variables. According to (1), let F and U denote the matrices of common and unique factors, respectively. Then, the sample analogue of the m-factor model (1) and the adopted assumptions imply that EFA represents Z as follows:

$$Z = F\Lambda^\top + U\Psi, \tag{3}$$

subject to $F^\top F = I_m$, $U^\top U = I_p$, $U^\top F = O_{p\times m}$ and Ψ diagonal. The EFA data representation (3) implies representation of the sample correlation matrix as:

$$Z^\top Z = \Lambda\Lambda^\top + \Psi^2 \quad . \tag{4}$$

The classical EFA finds Λ and Ψ by fitting the model $\Lambda\Lambda^\top + \Psi^2$ to $Z^\top Z$ with respect to some goodness-of-fit criterion, e.g. least squares (LS) or maximum likelihood. As Λ and Ψ have $pm + p$ (unknown) parameters and $Z^\top Z$ has $p(p+1)/2$ informative entries, the EFA problem becomes reasonable for certain values of m. Suppose that a pair $\{\Lambda, \Psi\}$ is already found by solving (4). The problem that many possible F exist for certain Z, Λ and Ψ is known as factor *indeterminacy* Mulaik (2005).

De Leeuw (2004); Trendafilov and Unkel (2011) recently proposed another view at the EFA model (1) and its sample version (3), by considering it as a specific data matrix decomposition. Then, the EFA problem is to minimize the following LS (least squares) criterion:

$$\Delta(F, \Lambda, U, \Psi) = ||Z - F\Lambda^\top - U\Psi||^2, \tag{5}$$

where the norm is the Frobenius matrix norm $\|A\|^2 = \sum_{i,j} a_{ij}^2$. As before, $F^\top F = I_m$, $U^\top U = I_p$, $U^\top F = O_{p\times m}$ and Ψ is square and diagonal. In (Trendafilov and Unkel, 2011), it is proved that this EFA reformulation gives unique Ψ and Λ (if chosen lower triangular) and quantitative characterization of the factor indeterminacy.

In this paper, we also consider EFA as a specific data matrix decomposition (5). We analyse in detail the factor indeterminacy using optimization on the Stiefel manifold. In particular, we show that *the classical EFA model is not well defined when* $n > m + p$. For years, the authors are looking for ways to extract well-defined factors by requiring them to possibly satisfy additional utility features. We believe that replacing the classical EFA model with a new well-defined one is a more reasonable mode of attack. In the new model, we approximate, in a least squares sense,

$$Z \approx F\Lambda^\top + U\Psi^\top, \quad F \in \mathbb{R}^{n\times m}, \quad \Lambda \in \mathbb{R}^{p\times m} \quad U \in \mathbb{R}^{n\times k}, \quad \Psi \in \mathbb{R}^{p\times k},$$

for $m + k < \min(n, p)$, with the constraint $(FU)^\top(FU) = I$. Ψ is no longer required to be diagonal, but the columns of Ψ are *sparse* and *structurally orthogonal*. The new model overcomes several pitfalls of the classical one. Its essence is that it allows for each "unique" factor u_i to contribute to more than one observable variable z_j. In this sense, u_i become shared factors, rather than unique as in the classical EFA. In addition, as k is the rank of the factor $U\Psi^\top$, it can be

seen as a parameter that determines the degree of dimension reduction or data compression of the model.

The paper is organized as follows. In Sect. 2, we provide a new characterization of the well-known factor indeterminacy problem and show that for $n > m + p$ the classical EFA model does not make sense. Section 3 introduces the alternative model and algorithm called for short the *semi-sparse* PCA (SSPCA),[1] which etymology becomes clear in Sect. 3.1. In the first step of the algorithm, the singular value decomposition (SVD) of the data matrix is computed. Then, an alternating procedure is executed: for fixed nonzero positions of Ψ, it finds U, satisfying the orthogonality constraint. This problem is a Procrustes problem that is solved using the SVD of a certain matrix. For fixed U, the objective function is minimized with respect to the nonzero positions of Ψ. In Sect. 4, we illustrate the new SSPCA algorithm on a couple of well-known large data sets traditionally used with other dimension reduction techniques. We also show that SSPCA is quite stable by running it on perturbed data with different levels of noise. Although the initial goal of this study is the development of a indeterminacy-free model as alternative to the classical EFA, the resulting data matrix decomposition is exceedingly resourceful. Section 5 demonstrates the versatility of the SSPCA method/algorithm which goes far beyond the standard EFA applications. Section 5.1 shows how SSPCA can perform sparse PCA and provide additionally the number of useful sparse PC for the particular data. The best existing sparse PCA techniques are generally faster, but lack this feature completely and work with prescribed number of PCs only. It is also shown in Sect. 5.2 that SSPCA can be very successful as a low-rank-plus-sparse matrix decomposition of large correlation matrices compared to a number of existing methods. We demonstrate that most of these approaches target exact fit to the data, which makes them vulnerable to either poor low rank or not sparse enough parts. A very recent approach (Aravkin et al., 2014) is able overcome this weakness. We demonstrate that SSPCA also solves the problem and moreover achieves better fit to the data. Finally, brief Conclusion summarizes the main points of the paper and potential directions for future work.

We let I_p denote the identity matrix of dimension p and define

$$I_{n,p} = \begin{pmatrix} I_p \\ 0 \end{pmatrix} \in \mathbb{R}^{n\times p},$$

where the columns of I_p are the canonical unit vectors e_j, for $j = 1, 2, \ldots, p$.

2. Why the Standard EFA does not Make Sense

The classical EFA model (4) with diagonal Ψ suffers from factor indeterminacy, i.e. there exist many possible F for certain Z, Λ and Ψ. As a result, a number of approaches were proposed to find optimal in some sense and thus unique, common factor scores F (Harman, 1976; Mulaik, 2010). Finding U is even more complicated. The major problems with the classic EFA model are summarized here:

Gramian Indeterminacy There is no unique Ψ^2, such that $Z^\top Z - \Psi^2$ is Gramian, i.e. it is a scalar product $(\Lambda\Lambda^\top)$ of a matrix;
Rotational Indeterminacy Apparently, for any orthogonal matrix Q, one has $Z^\top Z = \Lambda\Lambda^\top + \Psi^2 = \Lambda QQ^\top\Lambda^\top + \Psi^2$;

[1] The PCA concept is very specific and well defined: it is equivalent to a low-rank approximation of the data matrix using the singular value decomposition. Our proposed method is similar to PCA in a sense that it gives orthogonal factors, but it is not a PCA in the strict sense. In view of the fact that there are other related concepts, such as sparse PCA or robust PCA, which are not real PCA's, we decided to name our method semi-sparse PCA, SSPCA.

Factor Indeterminacy Even if Λ and Ψ^2 are unique, F and U are not. This was demonstrated by Guttman and Kestelman more than 50 years ago (Mulaik, 2005; Unkel and Trendafilov 2010). They define $F = Z(Z^\top Z)^{-1}\Lambda + SG$ and $U = Z(Z^\top Z)^{-1}\Psi - SG\Lambda^\top\Psi^{-1}$, where S is an arbitrary $n \times m$ matrix satisfying $S^\top S = I_m$ and $S^\top Z = 0_{m\times p}$, and G is a Gramian factor of $I_m - \Lambda^\top(Z^\top Z)^{-1}\Lambda$. The long history and evolution of the factor indeterminacy problem are carefully considered in (Steiger, 1979; Steiger and Schonemann, 1978).

In this section, we demonstrate the indeterminacy of the EFA problem in the standard case when $n \geq m + p$. Assume that $Z \in \mathbb{R}^{n\times p}$, with $n \geq m + p$, and $\text{rank}(Z) = p$. Consider the problem of minimizing (5) in slightly changed notations:

$$\min_{\widetilde{F},\widetilde{U},\Lambda,\Psi} \frac{1}{2}\left\| Z - \begin{pmatrix}\widetilde{F} & \widetilde{U}\end{pmatrix}\begin{pmatrix}\Lambda^\top \\ \Psi\end{pmatrix}\right\|^2, \tag{6}$$

where $\widetilde{F} \in \mathbb{R}^{n\times m}$ and $\widetilde{U} \in \mathbb{R}^{n\times p}$, under the constraints that $(\widetilde{F}\ \widetilde{U})^\top(\widetilde{F}\ \widetilde{U}) = I_{m+p}$, and that Ψ is diagonal. To this end, we first make an initial transformation of the problem.

Consider the QR decomposition of the $n \times p$ data matrix Z,

$$Z = Q\begin{pmatrix}R \\ 0\end{pmatrix}, \tag{7}$$

where $Q \in \mathbb{R}^{n\times n}$ is orthogonal and $R \in \mathbb{R}^{p\times p}$ is upper triangular and nonsingular (due to the full column rank of Z). Defining

$$X = \begin{pmatrix}R \\ 0\end{pmatrix}, \quad F = Q^\top\widetilde{F}, \quad U = Q^\top\widetilde{U}, \tag{8}$$

we see that (6) with the constraints is equivalent to

$$\min_{F,U,\Lambda,\Psi} \frac{1}{2}\left\| X - \begin{pmatrix}F & U\end{pmatrix}\begin{pmatrix}\Lambda^\top \\ \Psi\end{pmatrix}\right\|^2, \tag{9}$$

under the constraint

$$(F\ U)^\top(F\ U) = I_{m+p}. \tag{10}$$

The constraint on Ψ remains the same.

For any $(F\ U)$ satisfying (10), enlarge the matrix so that $(F\ U\ Y_\perp)$ is an orthogonal matrix. Then

$$\begin{aligned}\Theta &= \left\| X - (F\ U)\begin{pmatrix}\Lambda^\top \\ \Psi\end{pmatrix}\right\|^2 = \left\| \begin{pmatrix}F^\top \\ U^\top \\ Y_\perp^\top\end{pmatrix} X - \begin{pmatrix}I & 0 \\ 0 & I \\ 0 & 0\end{pmatrix}\begin{pmatrix}\Lambda^\top \\ \Psi\end{pmatrix}\right\|^2 \\ &= \|F^\top X - \Lambda^\top\|^2 + \|U^\top X - \Psi\|^2 + \|Y_\perp^\top X\|^2.\end{aligned}$$

After introducing the following partitioning

$$\begin{aligned}\begin{pmatrix}F & U\end{pmatrix} &= \begin{pmatrix}F_1 & U_1 \\ F_2 & U_2\end{pmatrix}, F_1 \in \mathbb{R}^{p\times m},\ U_1 \in \mathbb{R}^{p\times p}, \\ &F_2 \in \mathbb{R}^{(n-p)\times m},\ U_2 \in \mathbb{R}^{(n-p)\times p},\end{aligned} \tag{11}$$

we can write

$$\Theta = \|F_1^\top R - \Lambda^\top\|^2 + \|U_1^\top R - \Psi\|^2 + \|Y_\perp^\top X\|^2. \tag{12}$$

Since Λ is not constrained, the first term can be made equal to zero for any F. Ψ can be chosen arbitrarily as long as it is diagonal, and therefore, the minimization of the second term is equivalent to the minimization of the sum of squares of the off-diagonal elements of $U_1^\top R$, which, in turn, is equivalent to

$$\max \|\mathcal{D}(U_1^\top R)\|^2 = \max \sum_1^p \left(r_i^\top u_i^{(1)}\right)^2, \tag{13}$$

under the constraint $U^\top U = I_p$, and where r_i and $u_i^{(1)}$ are the column vectors of R and U_1, respectively. Note, that for $A \in \mathbb{R}^{p\times p}$ we define $\mathcal{D}(A) = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{pp})$ to be a diagonal matrix.

In order to analyse the maximization problem (13) with the constraint $U^\top U = I_p$, we will here consider it as an optimization problem where the unknown quantity U is required to lie on a Stiefel manifold. Manifold optimization for linear algebra problems is treated in (Absil et al., 2008; Edelman et al., 1998). Here it is enough to use the fact that in Stiefel optimization, essentially, the constraints are included in the solution geometry, and stationary points are not characterized using the standard gradient, but rather the Stiefel gradient.

Consider

$$\begin{aligned}\max_{U^\top U=I} \Gamma(U) &= \max_{U^\top U=I} \frac{1}{2}\|\mathcal{D}(U^\top X)\|^2 = \max_{U^\top U=I} \frac{1}{2}\|\mathcal{D}(U_1^\top R)\|^2 \\ &= \max_{U^\top U=I} \frac{1}{2}\sum_1^p \left(r_i^\top u_i^{(1)}\right)^2.\end{aligned} \tag{14}$$

The gradient of Γ in the ambient coordinate system is

$$\Gamma_U = X\,\mathcal{D}(U^\top X) = \begin{pmatrix} R \\ 0 \end{pmatrix} \mathcal{D}(U_1^\top R),$$

and, making it a Stiefel gradient (Edelman et al., 1998, Sect. 2.4.4),

$$\begin{aligned}\nabla\Gamma &= \Gamma_U - U\Gamma_U^\top U = X\,\mathcal{D}(U^\top X) - U\,\mathcal{D}(U^\top X)X^\top U \\ &= \begin{pmatrix} R \\ 0 \end{pmatrix} \mathcal{D}(U_1^\top R) - \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \mathcal{D}(U_1^\top R) R^\top U_1.\end{aligned} \tag{15}$$

At a stationary point the gradient is equal to zero, a fact that is used in the following lemma.

Lemma 2.1. *Assume that U is a stationary point that is the maximizer of* (14)*. Then*

$$U = \begin{pmatrix} U_1 \\ 0 \end{pmatrix}, \tag{16}$$

where $U_1 \in \mathbb{R}^{p\times p}$ is orthogonal.

The proof is given in 6. The result is quite natural, due to the fact that the data lie in a $p-$dimensional subspace of $\mathbb{R}^n$. Then one would expect that the optimum is attained when all the degrees of freedom of U are used to maximize the objective function inside that subspace.

Going back to (12), we now have the following result.

Theorem 2.2. *Assuming that $m + p = n$ the EFA problem (9), with constraints (10) and Ψ diagonal, determines U and the subspace spanned by the columns of F. In the case, $n > m + p$ U is determined, but F is undetermined. At the optimum, the parameter Λ always has the value 0.*

Proof. First let $m + p = n$. The objective function (12) is

$$\|F^\top X - \Lambda^\top\|^2 + \|U^\top X - \Psi\|^2,$$

and the first term can always be made equal to zero due to the fact that Λ is unconstrained (however, we will see below that its value is equal to zero at the optimum). The matrix $(F\ U)$ is square and orthogonal, and at the minimum for the second term it has the structure

$$(F\ U) = \begin{pmatrix} 0 & U_1 \\ F_2 & 0 \end{pmatrix},$$

due to Lemma 2.1. Thus, the subspace spanned by F is determined.

Next consider the case $n > m + p$, where the objective function is

$$\Theta = \|F^\top X - \Lambda^\top\|^2 + \|U^\top X - \Psi\|^2 + \|Y_\perp^\top X\|^2.$$

If we can show that at the minimizer of the second term, both the first and the third terms can be made equal to zero, then we can get the overall optimum by minimizing the second term separately.

Again, as Λ is unconstrained, the first term can always be made equal to zero. From Lemma 2.1, we see that at the minimum for the second term, the orthogonal matrix $(F\ U\ Y_\perp)$ must have the structure

$$(F\ U\ Y_\perp) = \begin{pmatrix} 0 & U_1 & 0 \\ F_2 & 0 & (Y_\perp)_2 \end{pmatrix},$$

where $F_2^\top (Y_\perp)_2 = 0$. Therefore, the third term is

$$\|Y_\perp^\top X\|^2 = (0\ (Y_\perp)_2^\top) \begin{pmatrix} R \\ 0 \end{pmatrix} = 0.$$

Thus, at the minimum of Θ neither F_2 nor $(Y_\perp)_2$ are determined by the model.

In both cases, Λ is equal to zero, since $F^\top X = 0$ at the optimum. □

The proof is completely based on the orthogonality properties of the columns of F and U (and $Y_\perp$) at the optimum. Those properties are the same also if we make the transformation back from (9) to the original problem (6). Therefore, the theorem holds also for (6).

In the next section, we propose an alternative EFA model, which is well defined. The analysis above shows that the standard EFA model suffers from indeterminacy. It is also easy to see another weakness. In data sets of today we can often expect some columns of Z to be close to linearly dependent. The model would associate such columns with different vectors u_j that are orthogonal; this would be unnatural and lead to numerical instabilities. The model described in the next section does not suffer from this deficiency.

3. New Model and Algorithm

A model and an optimization criterion that does not determine the model parameters and has a parameter that at the solution is always equal to zero, cannot be considered as well formulated. Theorem 2.2 shows that the classical EFA model is useless for $n \geq m + p$. The theorem also indicates that too many orthogonal vectors are included in the model. Based on this, we suggest an alternative model, where the total number of columns in F and U is less than $\min(n, p)$. In our new model, the "unique" factors, the u_j's, are not necessarily related to a single original variable. It will be shown that such a weakening of the standard EFA model assumptions makes its parameters identifiable.

3.1. Model

We start with the original problem with $Z \in \mathbb{R}^{n\times p}$, allowing for $n \geq p$ or $n < p$. Consider

$$\min_{\widetilde{F},\widetilde{U},\Lambda,\Psi} \frac{1}{2} \left\| Z - \begin{pmatrix} \widetilde{F} & \widetilde{U} \end{pmatrix} \begin{pmatrix} \Lambda^\top \\ \Psi^\top \end{pmatrix} \right\|^2,$$

where $\widetilde{F} \in \mathbb{R}^{n\times m}$, for some specified m.

Assume that the rank of Z is equal to s and let $Z = Q\Sigma V^\top$ be the SVD of the data matrix, where $Q \in \mathbb{R}^{n\times s}$, $\Sigma \in \mathbb{R}^{s\times s}$, and $V \in \mathbb{R}^{p\times s}$, and put $\mathbb{R}^{s\times p} \ni R = \Sigma V^\top$. The diagonal matrix Σ has full rank. Since the data column vectors are in the subspace spanned by the columns of Q, we also require $(\widetilde{F}\ \widetilde{U})$ to be in that subspace. We will choose $\widetilde{U} \in \mathbb{R}^{n\times k}$, for some $k \leq s - m$ (see Sect. 3.3). Then we put

$$(\widetilde{F}\ \widetilde{U}) = Q(F\ U), \qquad F \in \mathbb{R}^{s\times m}, \quad U \in \mathbb{R}^{s\times k}, \tag{17}$$

where the columns of $(F\ U)$ are orthonormal. Thus, we have the equivalent minimization problem,

$$\min_{F,U,\Lambda,\Psi} \left\| R - (F\ U) \begin{pmatrix} \Lambda^\top \\ \Psi^\top \end{pmatrix} \right\|, \qquad \Lambda \in \mathbb{R}^{p\times m}, \quad \Psi \in \mathbb{R}^{p\times k}. \tag{18}$$

In the (dysfunctional) EFA model of Sect. 2, each column vector r_j of R was associated with a unique vector u_j. We now relax the uniqueness requirement and allow several vectors in R to share the same u_j. This is natural, as we will be dealing with data matrices with collinearities, i.e. where there are column vectors that are (almost) linearly dependent. We prescribe each row of Ψ to have one nonzero element, denoted by ψ_j, and allow the columns of Ψ to have more than one nonzero element. This sparse construct was already used in (Adachi and Trendafilov, 2017) to achieve sparse factor loadings Λ. Note, that the proposed new model (18) has a dense Λ and sparse (nondiagonal) Ψ. With this prescription, the minimization problem (18) corresponds to modelling each column of R as

$$r_j \approx \sum_{i=1}^{m} \lambda_{ji} f_i + \psi_j u_{\mu(j)}, \qquad j = 1, 2, \ldots, p, \tag{19}$$

where $\mu(j)$ is a function that chooses one of the column vectors of U,

$$\mu : \{1, 2, \ldots, p\} \mapsto \{1, 2, \ldots, k\}.$$

Thus,

$$\Psi^\top = \left(\psi_1 e_{\mu(1)}\ \psi_2 e_{\mu(2)}\ \cdots\ \psi_p e_{\mu(p)}\right). \tag{20}$$

We start the computations for determining U and Ψ with k columns in the two matrices. Due to the fact that we allow several vectors from R in (19), say r_i and r_j, to share the same u_ν, i.e. $\mu(i) = \mu(j) = \nu$, it is quite likely that when the algorithm has run to finish, there will be zero columns in Ψ, which means that some of the u_ν's will not occur in any model (19). Such u_ν's will be called *passive*, and those that do occur in the model will be called *active*.

In matrix notation, (19) becomes $R \approx F\Lambda^\top + U\Psi^\top$, which gives

$$R^\top R \approx (F\Lambda^\top + U\Psi^\top)^\top (F\Lambda^\top + U\Psi^\top) = \Lambda\Lambda^\top + \Psi\Psi^\top. \tag{21}$$

The matrix $P = \Psi\Psi^\top$ cannot be diagonal,

$$p_{ij} = \begin{cases} \psi_i\psi_j, & \text{if } r_i \text{ and } r_j \text{ are associated with the same } u \text{ vector,} \\ 0, & \text{otherwise.} \end{cases}$$

However, $\Psi^\top\Psi$ is diagonal. Furthermore, the columns of Ψ are *structurally orthogonal*, in the sense that $|\Psi|^\top|\Psi|$ is diagonal, where $|\Psi|$ is the corresponding matrix of absolute values.

The new "unique" factors U are shared. The first term $\Lambda\Lambda^\top$ of the new decomposition (21), as its classical EFA analogue, takes care for the low-rank approximation of $R^\top R$. The second term $\Psi\Psi^\top$ provides sparse adjustment of the low-rank part, in order to improve the overall fit. The two parts of the new matrix factorization are orthogonal (uncorrelated) as in the classical EFA. They are spanned by vectors taken from orthogonal subspaces. In this sense, they should be clearly distinguished, in the same way "common" and "unique" factors are in the classical EFA. However, it is important to recognize that the unique factors are "unique" because Ψ is assumed diagonal, not because they come from subspace orthogonal to the one of the common factors. The purpose of the two orthogonal subspaces is to deliver a model as a sum of two terms, as in (4) and (21). If one allows for diagonal Ψ with few nonzero off-diagonal entries, then the classical EFA will change considerably. In this sense, in the new model, it seems more appropriate to speak about *adjusting* factors U, rather than for unique ones as in the classical EFA. Finally, there is no unique $\Psi\Psi^\top$, such that $Z^\top Z - \Psi\Psi^\top$ is Gramian, i.e. the Gramian indeterminacy remains.

The new model (19) can be written as $R \approx \Lambda^\top + U\Psi^\top = [F\ U][\Lambda\ \Psi]^\top$. This block matrix approximation is a kind of PCA-like factorization of R. For this reason, it can be called for short the semi-sparse PCA (SSPCA), because the block loading matrix is composed by dense Λ and sparse Ψ.

The interpretation of the new model SSPCA is changed with the size and format of the data analysed. For small data sets, the interpretation understandably concentrates on specific input variables. For a large number of variables, this strategy becomes impractical. This is reflected in the modern approach to produce sparse solutions (loadings) where the unimportant entries are simply made zeros (Trendafilov et al., 2017). It is worth mentioning here that in some modern large applications the interpretation is somewhat less important, what counts is how the model performs in terms of the application. For example, latent semantic indexing (LSI) is a widely used technique in text mining. LSI is a special form of PCA/SVD and its mathematical interpretation is clear. However, the application of that principle in LSI is not very informative (Eldén, 2007, 11.3). The new SSPCA could perhaps have a more natural interpretation. But that goes far beyond the aims of the present paper. The important bit is that the new method decomposes the sample

correlation/covariance matrix into a sum of low-rank approximating part and sparse adjustment part, which sparseness and size are self-formed and data driven.

3.2. *Algorithm*

Consider the residual (18). We will first assume that the function μ is given, and describe a method for solving the problem of minimizing the residual with respect to the unknowns $(F\ U)$, $\Lambda^\top$ and Ψ, under the constraints discussed above.

Recall that $\mathbb{R}^{s\times p} \ni R = \Sigma V^\top$, where the Σ and V are from the SVD of Z. Partition

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}, \quad V = (V_1\ V_2), \qquad \Sigma_1 \in \mathbb{R}^{m\times m}, \quad V_1 \in \mathbb{R}^{p\times m},$$

and put

$$R = \begin{pmatrix} R_1 \\ R_2 \end{pmatrix} := \begin{pmatrix} \Sigma_1 V_1^\top \\ \Sigma_2 V_2^\top \end{pmatrix}.$$

We propose to choose F and $\Lambda^\top$ so that $F\Lambda^\top$ is the best rank-m approximation of R. Due to the definition of R, we have

$$F = \begin{pmatrix} I_m \\ 0 \end{pmatrix}, \qquad \Lambda^\top = R_1 = \Sigma_1 V_1^\top, \tag{22}$$

which implies uniqueness of F and Λ. Moreover, (22) implies that Λ has *no* rotational freedom any more as in the classical EFA, i.e. it cannot be rotated without changing the model fit. Indeed, any counter rotation of F will destroy its special structure defined in (22). Thus, the rotational indeterminacy (of Λ) in the classical EFA is solved in the proposed new SSPCA model, as well as the common factor scores F indeterminacy.

Now, note that the best rank-m approximation of the original matrix Z is $QF\Sigma_1 V_1^\top =: Q_1\Sigma_1 V_1^\top$, where $Q_1 \in \mathbb{R}^{n\times m}$. The residual becomes

$$\left\| \begin{pmatrix} 0 \\ R_2 \end{pmatrix} - U\Psi^\top \right\|.$$

The requirement $F^\top U = 0$ implies

$$U = \begin{pmatrix} 0 \\ U_2 \end{pmatrix},$$

where $U_2 \in \mathbb{R}^{(s-m)\times k}$. As the columns of U are to be orthogonal and normalized, the same applies to U_2. Thus, we arrive at the minimization problem,

$$\min_{U_2,\Psi} \|R_2 - U_2\Psi^\top\|, \tag{23}$$

with the requirements that ${U_2}^\top U_2 = I$, and that Ψ has the structure (20), i.e. the columns of Ψ are structurally orthogonal.

If Ψ is fixed in the minimization problem (23), then we have an orthogonal Procrustes problem, which can be solved using the SVD of $R_2\Psi$ (Golub and Van Loan, 2013, Chapter 12). Let

$$R_2\Psi = PSW^\top, \quad P \in \mathbb{R}^{(s-m)\times k}, \quad S \in \mathbb{R}^{k\times k}, \quad W \in \mathbb{R}^{k\times k},$$

where S is diagonal be the thin SVD. The solution of (23) is

$$U_2 = PW^\top, \tag{24}$$

which is unique, for a fixed Ψ, as solution of a standard Procrustes problem (SVD).

The overall SSPCA algorithm will be alternating: given μ and Ψ, compute U_2 from (24). Then, given U_2, update μ and Ψ.

The update of the function μ and Ψ, given U_2, is made based on the observation that the solution of the least squares problem

$$\min_{\psi_j} \|r_j - \psi_j u_\nu\|, \qquad 1 \le \nu \le k, \tag{25}$$

is $\psi_j = r_j^\top u_\nu$, and the residual is $\|(I - u_\nu u_\nu^\top)r_j\|$. If the smallest residual is obtained for $\nu = i$, then $\mu(j) := i$.

As with any iterative method, the choice of the starting point is important. Random starts are always an option, but having a good rational start is preferable. For this reason, we note that $R_2 = \Sigma_2 V_2^\top = I_{s-m}\Sigma_2 V_2^\top$, i.e. the left singular vectors of R_2 are the canonical unit vectors $e_i \in \mathbb{R}^{s-m}$. Thus, those vectors are the "best" basis vectors for the column space of R_2. Therefore, it is natural to choose the identity matrix $I_{s-m,k}$ as a starting approximation for U_2 in the above alternating procedure.

The number of active columns may vary during the optimization procedure. When the optimal solution is computed, we remove the passive columns in U_2 and delete the corresponding columns of Ψ.

The SSPCA procedure is summarized in Algorithm 1, where MATLAB-like notation is used.

Algorithm 1
Given R_2, compute U_2, Ψ, and μ.

{Initialization}
$U_2 := I_{s-m,k}$
for $j = 1 : p$ **do**
  $i := \max(\text{abs}(R_2(:, j)))$
  $\Psi(j, i) := R_2(i, j)$
  $\mu(j) := i$
end for
{Alternating procedure}
repeat
  Update U_2 from (24)
  Update μ and Ψ
until convergence
Remove all passive columns of U_2 and corresponding zero columns of Ψ

Let U_2 and Ψ denote the output of Algorithm 1, and put $\widetilde{U} = Q_2 U_2$. The algorithm gives an approximation of the original matrix

$$Z \approx Q_1 \Sigma_1 V_1^\top + \widetilde{U}\Psi^\top = \sum_{i=1}^{m} \sigma_i u_i v_i^\top + \sum_{i=1}^{k} \tilde{u}_i \phi_i^\top = \sum_{i=1}^{m} \sigma_i \tilde{f}_i v_i^\top + \sum_{i=1}^{k} \tilde{u}_i \phi_i^\top,$$

where the columns of Ψ are denoted ϕ_i. Thus, the first term is the best rank-m approximation, i.e. PCA. The second term has a left factor with orthonormal columns; the right factor is sparse with structurally orthogonal columns.

3.3. *Model Size*

We will refer to the number of columns in $(F\ U)$ (in the notation above $m + k$) as the *model size*. In our formulation of the SSPCA algorithm above, we assume that m and k are chosen a priori. After the optimal solution has been computed, it is essential to be able to test whether an increase in the dimension of U_2 will reduce the residual substantially. Note that one should not just take any vector orthogonal the active u_ν's, because such a vector would not be likely to become active in the subsequent optimization. Instead, we suggest the following choice, which is the closest vector not accounted for in the present model.

Choose W such that $(U_2\ W)$ is an orthogonal matrix, make the Ansatz $u = Wc$, where $c^\top c = 1$, and consider the residual for the enlarged model

$$\left\| R_2 - (U_2\ u) \begin{pmatrix} \Psi^\top \\ \theta^\top \end{pmatrix} \right\|^2 = \|U_2^\top R_2 - \Psi^\top\|^2 + \|W^\top R_2 - c\theta^\top\|^2.$$

Clearly the vector c that gives the best approximation is the first left singular vector of $W^\top R_2$. This enlarged model can then be taken as first approximation for a new run of the alternating algorithm.

4. Numerical Experiments

In all examples, we start with a data matrix Z that is centred and normalized to have columns of Euclidean norm 1. The computations were performed using MATLAB R2016b on a standard (2015) quad-core laptop computer under Ubuntu Linux. The codes used and the data matrices are available at http://users.mai.liu.se/larel04/EFA/.

Let

$$\delta = \frac{\epsilon}{\rho},$$

where ϵ is the maximum change of any element in Ψ between two iterations and ρ is the element of the largest magnitude in Ψ. The iterations were stopped when δ was less than 0.05.

4.1. *Complexity*

As the SSPCA algorithm is iterative, where the number of iterations is highly problem dependent, it is impossible to give detailed operation counts for the algorithm. Such counts would also be difficult to interpret due to the complexity of the computer architecture and of the programming environment. We will give here some timings below, based on the use of the MATLAB profiler, that indicate how large problems can be solved on a standard computer. In addition, we will see that the structure of a modern language like MATLAB, with highly optimized standard functions

(in our context the `svd` function, which by default is implemented with parallel execution on more than one core on the computer that we used), makes it very difficult or impossible to draw conclusions based on operation counts.

The SSPCA algorithm starts with the computation of the SVD of Z, so the total complexity exceeds that of PCA (i.e. the computation of an SVD). In every iteration, the evaluation of the new iterate requires the computation of the thin SVD of a matrix of dimension $(s-m)\times k$. In many applications, $s-m$ and k will be much smaller than n and p; therefore, for very large problems, one can expect the cost for computing the initial SVD to be dominating. That was actually the case in the example with cancer data below, where each SVD during the iterations required about 3% of the time for the initial SVD.

On the other hand, in our numerical experiments we saw that the determination of $\Psi^\top$, given U_2, took much longer time than the SVD. This computation which involves solving a number of simple least squares problems (25) has an operation count $\mathcal{O}(skp)$, which is comparable to that for the SVD's, but as it is coded in MATLAB, it is not optimized.

4.2. Real Data Experiments

Here we demonstrate the behaviour of the SSPCA algorithm on real data. We also perform simulation study to evaluate the effectiveness of the new method with respect to the level of data contamination. First, we find a solution of the particular data set and display record of the convergence history. This solution $\{F_0, U_0, \Lambda_0, \Psi_0\}$ is considered target in the consequent simulation study. Then, we contaminate the data by adding Gaussian noise as follows: $X := X + \sigma N$, where X is already standardized, and each element of σN is a normally distributed random number with mean 0 and standard deviation σ, which controls the level of contamination. However, depending on the particular number of observations, the standardized data will be in different numeric ranges. Thus, a level of contamination for one data set will have different effect for a different data set. In order to have a rough guidance for the appropriate level of contamination (not too small or too large), we assume that the magnitude of a "typical" value in a particular data set is $1/\sqrt{n}$. For example, the Cancer data set (Sect. 4.2.1) has $n = 216$ observations, and the magnitude of its "typical" value is 0.0680, while the MLL data set (Sect. 4.2.2) has $n = 72$ observations, and its "typical" magnitude is 0.1179. Thus, it is likely that a rather moderate contamination for the MLL data may have considerable contaminating effect for the Cancer data.

The SSPCA algorithm is applied to different *standardized* contaminated data, and the solutions $\{F, U, \Lambda, \Psi\}$ are compared to the target $\{F_0, U_0, \Lambda_0, \Psi_0\}$ (original solution) in the following way. We use as a discrepancy measure between two $m\times n$ matrices A_1 and A_2 their mean absolute error (MAE), i.e. $\|A_1 - A_2\|_{L_1}/mn$. To compare fairly Λ to Λ_0, we, in fact, solve a Procrustes problem $\|\Lambda Q - \Lambda_0\|_F$ and measure the MAE between ΛQ and Λ_0. Similarly, we compare the MAE of the rotated FQ to F_0. For U and Ψ, we only use their element-wise magnitudes, i.e. $\||\Psi| - |\Psi_0|\|_1$ and $\||U| - |U_0|\|_{L_1}$, where $|\ |$ denotes taking element-wise absolute value. This is because U cannot be rotated, as counter rotation of Ψ would destroy its sparse structure. Finally, we compute MAE of the normalized original fit $\|X - F_0\Lambda_0^\top - U_0\Psi_0^\top\|_F/\|X\|_F$ and the one for the perturbed data. To evaluate the perturbation influence on the number of location changes in Ψ, we also report $\sharp(\Psi)/\sharp(\Psi_0)$. The values smaller than 1 indicate that the original data need more location changes to reach the optimal Ψ than the perturbed one.

These same type of comparisons will be reported for each of the following two large data sets.

4.2.1. Cancer Data To test the algorithm for a medium size data set, we ran the algorithm with $m = 2$ on the ovarian cancer data from MATLAB's Statistics and Machine Learning Toolbox.[2]

[2] https://se.mathworks.com/help/stats/select-data-and-validation-for-classification-problem.html?s_tid=srchtitle.

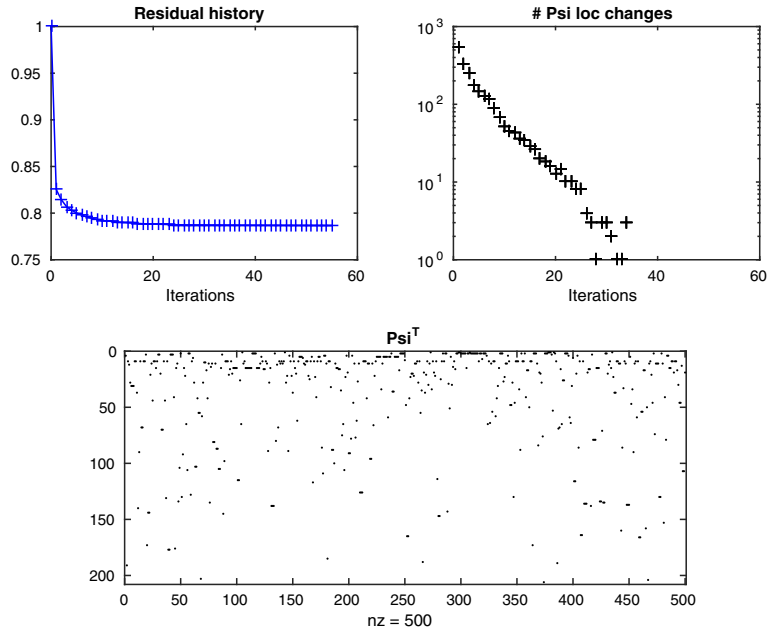

FIGURE 1.
Cancer data. Top left: Residual history ($\|R_2 - U_2\Psi^\top\|^2/\|R_2\|^2$) as a function of iterations. Top right: The number of element locations changed in Ψ (the last 21 iterations the nonzero locations remained fixed). Bottom: Plot of the nonzero structure of the matrix $\Psi^\top$. Each dot represents a nonzero element. For visibility, we only plot the first 500 columns of $\Psi^\top$.

The data matrix is 216×4000. We started with 214 columns in U_2; at the end 208 were active. The results are illustrated in Fig. 1. The relative residual after the PCA step was 0.831, and after the second stage

$$\left\| R - (F\ U) \begin{pmatrix} \Lambda^\top \\ \Psi^\top \end{pmatrix} \right\| / \|R\| = 0.737.$$

71 iterations were required. The time for the initial SVD was around 0.6 s, and the total execution time was about 150 s. Using the MATLAB profiler, it was seen that more than 92% of this time was used to update Ψ. There were in total 2481 location changes in the entries of Ψ until the optimal ones were found. This can be explained by the fact that, while the SVD function is highly optimized, the computation of Ψ is to a high degree coded in the MATLAB language.

In the following Table 1, we report the resulting MAEs and standard deviations (SDs) from 10 runs of the algorithm with contaminated data for each value of σ. The important observation is that Λ, F and U are very stable, with F and U hardly changing with the increase in the noise. The changes in the objective function become considerable for $\sigma > .05$, while Ψ is very unstable. All perturbed data require less location changes in Ψ, than those for the original data, occasionally up to 30%. If one assumes that the "typical" value in the standardized data is $\pm 1/\sqrt{n} = \pm 1/\sqrt{216} = \pm 0.0680$, adding noise of $\pm .05$ already means significant contamination. This suggests that the algorithm is pretty stable.

4.2.2. MLL Data The MLL data contain 72 leukaemia samples of which 24 come from acute lymphoblastic leukaemia (ALL), 20—from mixed lineage leukaemia (MLL), and 28—from acute

TABLE 1.
MAE/SD for contaminated Cancer data.

σ	Λ	Ψ	F	U	Fit	Ψ Loc change
.01	0.0081/0.0001	0.1804/0.0257	0.0004/0.0000	0.0022/0.0000	0.0048/0.0008	0.9752/0.0593
.03	0.0326/0.0002	0.2361/0.0506	0.0011/0.0000	0.0022/0.0000	0.0402/0.0014	0.9576/0.0900
.05	0.0514/0.0003	0.2849/0.0373	0.0021/0.0001	0.0023/0.0000	0.0831/0.0015	0.9658/0.1293
.07	0.0594/0.0005	0.2845/0.0233	0.0032/0.0001	0.0023/0.0000	0.1199/0.0009	0.7960/0.0843
.10	0.0718/0.0005	0.2565/0.0069	0.0045/0.0002	0.0023/0.0000	0.1545/0.0006	0.6971/0.0498
.15	0.0815/0.0008	0.2830/0.0051	0.0069/0.0003	0.0023/0.0000	0.1809/0.0004	0.7410/0.0201
.20	0.0776/0.0005	0.2938/0.0018	0.0100/0.0005	0.0023/0.0000	0.1914/0.0003	0.7680/0.0169

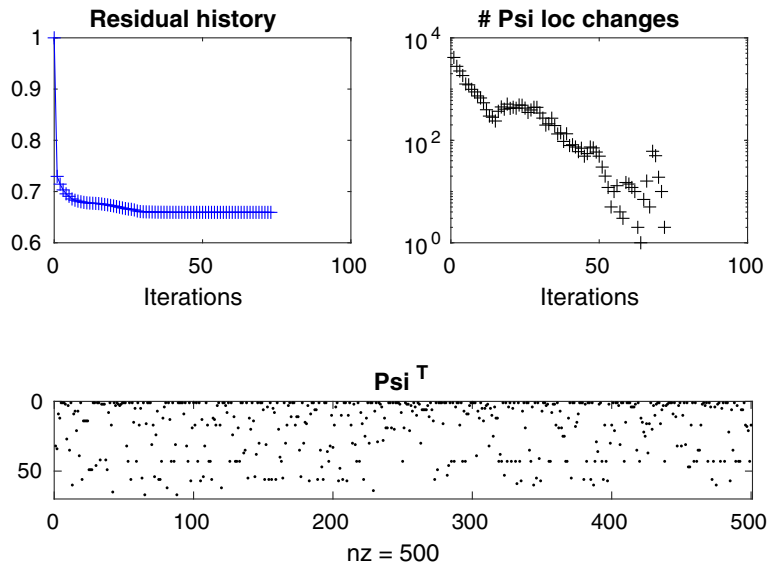

FIGURE 2.
MLL data. Top left: Residual history. Top right: The number of element locations changed in Ψ. Bottom: Nonzero structure of the matrix $\Psi^\top$, first 500 columns.

myeloid leukaemia (AML). The data matrix[3] has dimension 72×12582 and has been studied first by Armstrong et al. (2002).

We used $m = 2$, and started with 72 columns in U_2; after the problem was run to completion 69 were active. The relative residual was 0.847 after the PCA step and 0.732 after the complete run. In total, 95 iterations were performed. The time for the initial SVD was about 60 s, and the total execution time was about 187 s. Using the MATLAB profiler, it was seen that more than 95% of this time was used to update Ψ. There were in total 14811 location changes in the entries of Ψ until the optimal ones were found. The results are illustrated in Fig. 2.

Now, we investigate the sensitivity of the algorithm to noisy data. In the following Table 2, we report the MAEs and SDs from 10 runs of the algorithm with contaminated data for each value of σ. The results resemble our findings for the Cancer data in Sect. 4.2.1. Again, Λ, F and U hardly change with the increase in the noise. The changes in the objective function become

[3] Available from http://datam.i2r.a-star.edu.sg/datasets/krbd/Leukemia/MLL.html.

TABLE 2.
MAE/SD for contaminated MLL data.

σ	Λ	Ψ	F	U	Fit	Ψ Loc change
.01	0.0074/0.0000	0.1722/0.0131	0.0002/0.0000	0.0003/0.0000	0.0019/0.0001	0.9540/0.0389
.03	0.0241/0.0001	0.1839/0.0298	0.0007/0.0000	0.0003/0.0000	0.0152/0.0002	0.9897/0.0434
.05	0.0431/0.0002	0.2380/0.0124	0.0013/0.0001	0.0003/0.0000	0.0379/0.0003	1.0866/0.0571
.07	0.0599/0.0002	0.2253/0.0146	0.0019/0.0001	0.0003/0.0000	0.0634/0.0003	1.2023/0.0547
.10	0.0821/0.0004	0.2632/0.0091	0.0033/0.0002	0.0004/0.0000	0.0993/0.0004	1.1665/0.0828
.15	0.1026/0.0004	0.2884/0.0032	0.0041/0.0003	0.0004/0.0000	0.1386/0.0004	0.9239/0.0543
.20	0.1124/0.0005	0.3078/0.0011	0.0056/0.0004	0.0004/0.0000	0.1585/0.0002	0.9386/0.0398

considerable for $\sigma > .1$. As before, Ψ is very unstable. However, the number of location changes in Ψ for the perturbed data does not deviate much from those for the original data (except for two levels of σ).

Assuming that the "typical" value in the standardized data is $\pm 1/\sqrt{n} = \pm 1/\sqrt{72} = \pm 0.1179$, it shows again that the algorithm copes well with noise of up to $\pm .07 (\pm .1)$.

5. Comparison with Other Related Data Matrix Factorizations

Historically, PCA is the most popular dimension reduction technique. PCA of a given $n \times p$ data matrix X is obtained by a truncated SVD of X. For a specified m, the best LS approximation to X of rank m is given by $X \approx QSP^\top$, where S is $m \times m$ diagonal, containing the m largest singular values of X, and Q and P which are, respectively, $n \times m$ and $p \times m$ orthonormal, containing the corresponding singular vectors. Then, why look for other data matrix factorizations, which are suboptimal compared to PCA/SVD?

There are several reasons, but probably the most important one is that the modern data are very large and their PCA solutions are difficult to interpret. In the last couple of decades, there appeared a great number of alternative dimension reduction techniques producing sparse solutions, which are easier to interpret. One big group of methods modify PCA to produce sparse component loadings P and are known now as sparse PCA/SVD (Jolliffe et al., 2003; Journée et al., 2010; Shen and Huang, 2008; Witten et al., 2009). Another type of methods/algorithms aims at low-rank-plus-sparse matrix decomposition of large data containing noise and/or missing entries, problems known also as matrix completion (Cai et al., 2008; Candès et al., 2009).

We will show now that the proposed new SSPCA model and algorithm can, in fact, serve for both of the above purposes. The SSPCA model $Z \approx F\Lambda^\top + U\Psi^\top$ incorporates the two extreme cases $Z \approx F\Lambda^\top$ and $Z \approx U\Psi^\top$, which themselves correspond to the standard PCA/SVD of Z and to its sparse PCA/SVD, respectively. The "intermediate" SSPCA models can be seen as low-rank-plus-sparse matrix decompositions of the sample correlation matrix of the data.

5.1. *Sparse PCA*

Indeed, the proposed SSPCA algorithm turns into a sparse PCA by taking $m = 0$, i.e. transforming (5) into $Z \approx U\Psi^\top$. This can be particularly convenient if you have to analyse a sparse matrix. Then, each data row is modelled by a linear combination (not sparse) of sparse rows.

To explore this option, we consider data from the Medline database containing biomedical abstracts. This particular data matrix has dimension 1033×4163 and was constructed from 1033

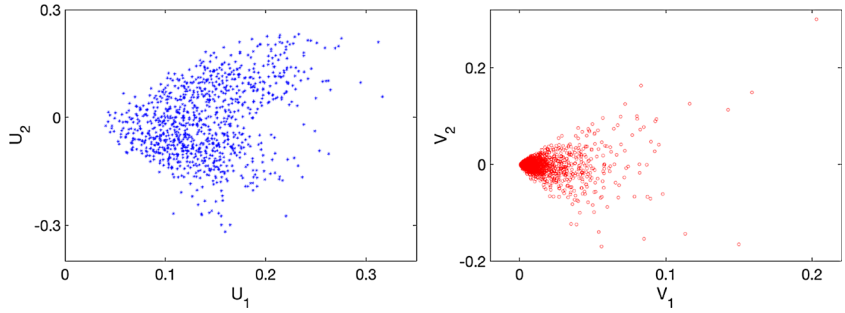

FIGURE 3.
PCA of Medline data. Left: First two-component scores. Right: First two-component loadings.

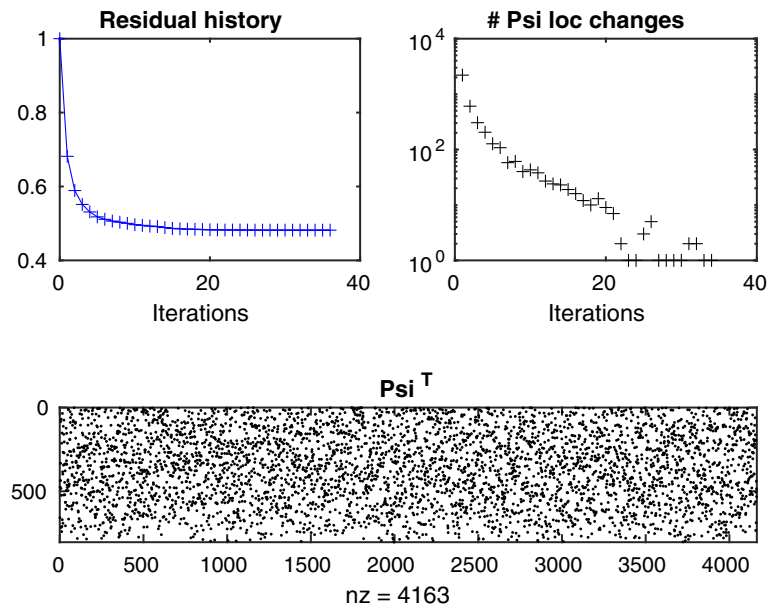

FIGURE 4.
Medline data. Top left: Residual history. Top right: Number of element locations changed in Ψ. Bottom: Nonzero structure of $\Psi^\top$.

abstracts, where stopwords were removed and stemming performed. For details, see (Eldén 2007, Chapter 11). The standard PCA/SVD solution is depicted in Fig. 3. One can hardly find any interesting pattern in the cloud of thousands points, except its fan shape. It is interesting can one get any more specific information by applying sparse PCA.

This example is also included to investigate additionally (to the examples in Sect. 4.2) the performance of the iterations for a matrix U_2 of relatively large dimension. In this case, we did not centralize or normalize the matrix. We used $m = 0$, and started with 1033 columns in U_2; after the problem was run 36 iterations to completion 803 were active. The relative residual was 0.694 after the complete run. The time for the initial SVD was around 3.4 s, and the total execution time was about 690 s. This information is summarized in Fig. 4. More than 93% of this time was used to update Ψ. When we ran the same problem starting with 300 columns in U_2, the execution time was about 187 s.

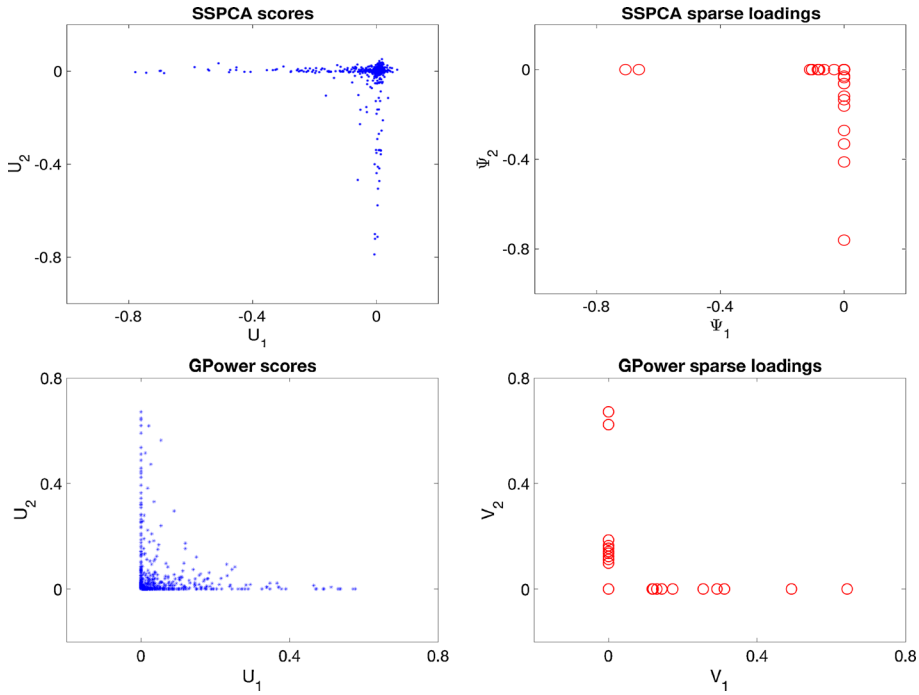

FIGURE 5.
Medline data. Top left: First two SSPCA scores. Top Right: First two SSPCA loadings. Bottom left: First two GPower scores. Bottom light: First two GPower sparse loadings.

TABLE 3.
Nonzero loadings for Medline data.

Method	Component	Nonzero locations
SSPCA	I	86 1124 1157 1616 1743 1797 1803 2810 3082 3340
	II	360 1083 1746 1884 2768 2819 3132 3493 3621 3699 3861
GPower	I	1157 1616 1704 1743 1750 1797 1965 2810 3098 3385
	II	442 553 850 2031 2083 2211 2212 2269 3464 3817

The top panel of Fig. 5 contains the plots of the first two columns of U and Ψ, which are regarded as score and sparse loadings matrices. The first two columns of Ψ contain 21 nonzero loadings. For comparison, we apply GPower (Journée et al., 2010) to the Medline data, which is considered the best sparse PCA algorithm. We produce two sparse components with 20 nonzero loadings (the closest we can get to 21). The bottom panel of Fig. 5 contains the plots of the first two sparse components found by GPower. Note, that Ψ is normalized, such that $\Psi^\top\Psi = I_{803}$ (and U is accordingly adjusted), because GPower produces orthonormal sparse loadings. The locations of the nonzero entries in both solutions are given in Table 3.

Both sparse solutions are very similar. The proposed SSPCA procedure needs 368 s (this experiment is with different computer) to find 4163×803 sparse matrix Ψ containing, of course, 4163 nonzero entries. The first two columns, with the largest column sums of squares, are displayed and discussed above. Now, we ask GPower to find 4163×803 sparse component loadings matrix from the original 1033×4163 data matrix. We use the same thresholding parameters as for the

two-component solution shown above (with 20 nonzeros). This takes 75 s; however, the resulting sparse matrix has 12674 nonzero entries, i.e. three times more than in our Ψ. Looking for a sparser matrix is generally more demanding, but it is not very reasonable to compare here the speed of the two algorithms, as their strengths and goals are different. If one simply needs only few most representative sparse components, then the right action would be to apply GPower (or other sparse PCA) for fast result. However, if one needs the whole matrix decomposition with its factors, then SSPCA is probably to be preferred as it finds the correct, active, inner size of the involved parameter (factoring) matrices. Finding very few sparse components may not be a satisfactory solution for many large data sets. The problem is that the spectrum of many large data is smoothly, slowly and gradually decreasing and does not provide clear-cut location to be used for dimension reduction. For example, the first two principal components of the Medline data explain less than 1% of the total variance. One needs 345 components to explain 50%. The SSPCA algorithm finds that the optimal number of components to use for the Medline data is 803, which explains 88% of the total variance.

5.2. *Low-Rank-Plus-Sparse Matrix Decomposition*

Suppose that the SSPCA algorithm was run and the optimal fit to the data Z is found to be $\{F, U, \Lambda, \Psi\}$. Then, the new SSPCA model definition (19) and the related constraints make it possible to find that $Z^\top Z \approx \Lambda\Lambda^\top + \Psi\Psi^\top$, where $Z^\top Z$ is the sample correlation matrix for standardized Z. In other words, our SSPCA algorithm can be seen as a kind of low-rank-plus-sparse matrix decomposition of the sample correlation matrix, because $\Lambda\Lambda^\top$ is a low-rank matrix and $\Psi\Psi^\top$ is sparse. By the way, the classical EFA can be seen as the oldest low-rank-plus-sparse matrix decomposition, because it looks for $\{\Lambda, \Psi\}$ and Ψ diagonal, such that $Z^\top Z \approx \Lambda\Lambda^\top + \Psi^2$.

Medline data are sparse and not convenient for analysis through correlations. Instead, we consider the Cancer data from Sect. 4.2.1, with sample correlation matrix $Z^\top Z$ of size 4000×4000. We want to compare the proposed SSPCA method/algorithm with methods for low-rank-plus-sparse matrix decomposition, also known as robust PCA (RPCA) (Candès et al., 2009). Our solution, obtained in Sect. 4.2.1 is used to form $\Lambda\Lambda^\top + \Psi\Psi^\top$, which gives a rank-two-plus-sparse matrix decomposition of the sample correlation matrix $Z^\top Z$. This solution is obtained for 69 s. The sparse term $\Psi\Psi^\top$ has 722842 nonzero entries (cardinality), which means that 4.52% of the entries of the sparse term are nonzeros. The achieved normalized fit $\|Z^\top Z - \Lambda\Lambda^\top - \Psi\Psi^\top\|_F / \|Z^\top Z\|_F$ is 0.2636.

A useful review of different approaches and algorithms for RPCA is given in (Candès et al., 2009, Sect. 5). First, we tried two algorithms (`LRSD` and `RobustPCA`) admitting very compact implementation of the alternating directions method (Yuan and Yang, 2013). However, they need considerable CPU time to decompose $Z^\top Z$ into a rank-two or rank-three matrix plus a 'sparse' part, which is, in fact, pretty dense. Thus, they are not taken for comparison. Next, we tried `partial-proximal-gradient-RPCA` based on the accelerated proximal gradient (Lin et al., 2009a). With $\lambda = 0.0009$, it decomposed $Z^\top Z$ for 105 s into a rank-two matrix plus a 'sparse' part with only 454 zero entries (out of 16×10^6), i.e. practically a dense matrix. Similar result was obtained by `inexact ALM RPCA` (Lin et al., 2009b) with $\lambda = 0.0009$, which is based on the augmented Lagrange multiplier method. In this case, $Z^\top Z$ was decomposed for 39 s into a rank-two matrix plus a 'sparse' part with cardinality 99.86%, i.e. again nearly dense.

Apparently, these methods compensate the lower rank with lower sparseness and vice versa. This is because they require that the low-rank part plus the sparse one should reconstruct the data *perfectly*. An algorithm that can maintain these two features separately is `fastRPCA` (Aravkin et al., 2014) and is available from Mathworks File Exchange. Of course, the exact fit to the input data is sacrificed. Again, we try to find a solution resembling our, i.e. decomposing $Z^\top Z$ into a rank-two matrix plus a sparse matrix with about 4.52% nonzero entries. Such a solution

is obtained with `solver-RPCA-Lagrangian` with parameters $\lambda_L = 145$ and $\lambda_S = .185$. Indeed, `fastRPCA` is much faster and produces such a solution (rank-two-plus-4.4%-sparse matrix) for less than 4 s. However, the normalized fit is 0.3283, which is 20% worse than the one achieved by SSPCA. It is unlikely that this is only due to the 0.12% difference in sparseness. This experiment suggests/indicates that the SSPCA can be a reasonably competitive method for RPCA of a large correlation matrix able to maintain simultaneously low-rank and sparse parts.

6. Conclusions

The purpose of this paper has been to reconsider the classical EFA and elevate it to a modern dimension reduction technique capable of producing well-defined parameter estimations.

We approach EFA as a specific matrix factorization problem. First, we reconsider the factor indeterminacy problem of the classical EFA model and reveal the reason for its numerical behaviour. To address this long-standing problem of EFA, we propose a new and more general SSPCA model, which permits factors to relate to more than one variable. Next, we develop a fast and stable algorithm for its solution applicable to any type of data (with more or less variables than observations). In addition, the construction of the new SSPCA algorithm leads to unique factor loadings and scores and thus circumvents the notorious EFA rotation indeterminacy.

We demonstrate that the new SSPCA model and algorithm can play valuable role among the existing methods for sparse PCA of large sparse data and robust PCA of large correlation matrices. The comparisons with other related methods show that the proposed EFA deserves further systematic work.

It is well known that alternating algorithms may have quite slow convergence. However, when SSPCA is applied to medium–large problems, the number of iterations was not excessive. We are currently working on the improvement in the algorithm to make it possible to use it for large and structured problems.

Proof of Lemma 2.1

Proof. In the proof we will, for convenience, denote $\mathcal{D} = \mathcal{D}(U_1^\top R)$. Without loss of generality, we assume that the diagonal elements are ordered, $|u_1^\top r_1| \geq |u_2^\top r_2| \geq \cdots \geq |u_p^\top r_p|$. Here r_i and u_i denote the i'th column of R and U_1, respectively.

Let the CS decomposition (Golub and Van Loan, 2013, Sect. 2.5.4) of U be

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix} \begin{pmatrix} C \\ S \end{pmatrix} V^\top, \quad C^2 + S^2 = I_p,$$

where Q_1, Q_2 and V are orthogonal and C and S are diagonal. The diagonal elements of C satisfy $c_1 \geq c_2 \geq \cdots \geq c_p \geq 0$. Clearly, the statement of the lemma is equivalent to $C = I_p$ and $S = 0$.

We insert the CS decomposition in (15), set $\nabla\Gamma = 0$ and multiply by

$$\begin{pmatrix} Q_1^\top & 0 \\ 0 & Q_2^\top \end{pmatrix}$$

from the left and by V from the right. We get

$$\begin{pmatrix} Q_1^\top R\mathcal{D}V \\ 0 \end{pmatrix} = \begin{pmatrix} CV^\top \mathcal{D}R^\top Q_1 C \\ SV^\top \mathcal{D}R^\top Q_1 C \end{pmatrix}.$$

With $B = Q_1^\top R\mathcal{D}V \in \mathbb{R}^{p\times p}$, we thus have $B = CB^\top C$, or, equivalently,

$$B = C^2BC^2. \tag{26}$$

Clearly, for any $b_{ij} \neq 0$ we must have $c_i = c_j = 1$. Assume that

$$B = \begin{pmatrix} B_s & 0 \\ 0 & 0 \end{pmatrix}, \quad B_s \in \mathbb{R}^{s\times s}, \tag{27}$$

and that, for some $1 \leq i \leq p$, $b_{is} \neq 0$, or, for some $1 \leq j \leq p$, $b_{sj} \neq 0$ (we allow $s = p$, in which case $B_s = B$). Then, since $b_{is} = c_i^2 b_{is} c_s^2$, or $b_{sj} = c_s^2 b_{sj} c_j^2$, and since the c_i's are ordered, we must have $c_1 = \cdots = c_s = 1$. Thus, if $s = p$, then $C = I_p$, and the lemma is true.
If $s < p$, C has the structure

$$C = \begin{pmatrix} I_s & 0 \\ 0 & C_2 \end{pmatrix},$$

where $C_2 = 0$ or its diagonal elements are nonnegative.

We now assume that

$$C_2 = \operatorname{diag}(c_{s+1}, \ldots, c_t, 0, \ldots, 0) \tag{28}$$

with $c_t > 0$, i.e. U_1 has rank $t < p$. We will show that then the stationary point does not correspond to a global maximum.

Due to (27), we can write

$$BV^\top = Q_1^\top R\mathcal{D} = \begin{pmatrix} \bar{B}_s & \bar{B}_{1s} \\ 0 & 0 \end{pmatrix}.$$

Consider the last row of $BV^\top$:

$$\begin{aligned} 0 &= \begin{pmatrix} q_p^\top r_1 & q_p^\top r_2 & \cdots & q_p^\top r_p \end{pmatrix} \begin{pmatrix} u_1^\top r_1 & & & \\ & u_2^\top r_2 & & \\ & & \ddots & \\ & & & u_p^\top r_p \end{pmatrix} \\ &=: c^\top \mathcal{D}, \end{aligned}$$

where r_i denotes the i'th column of R. Since R is nonsingular, there must exist at least one nonzero element in $c^\top$, say $q_p^\top r_k$. Then the corresponding element i $\mathcal{D}$ must be equal to zero, $u_k^\top r_k = 0$.

Under the assumption (28), U_1 has rank t: using the CS decomposition we can write $u_j = \sum_{i=1}^t c_i v_{ji} q_i$, for $j = 1, 2, \ldots, p$. Clearly q_p is orthogonal to $\{u_1, u_2, \ldots, u_p\}$. Thus, we can replace the column u_k in U_1 by q_p and make the objective function larger. It follows that the assumption that $\operatorname{rank}(U_1) = t < p$ cannot be valid at the global maximum.

It remains to consider the case when all the diagonal elements of C are positive, and U_1 is nonsingular. Due to the structure (27), we have

$$(Q_1^\top R)^{-1} B = \mathcal{D}V = \begin{pmatrix} \tilde{B}_s & 0 \\ \tilde{B}_{s1} & 0 \end{pmatrix}.$$

With the corresponding blocking $V = (V_1\ V_2)$, we have $\mathcal{D}V_2 = 0$, i.e. $\mathcal{D}$ has a null space of dimension $p - s$. Since the diagonal elements of $\mathcal{D}$ are ordered, it follows that $\mathcal{D}$ has the structure

$$\mathcal{D} = \begin{pmatrix} \mathcal{D}_s & 0 \\ 0 & 0 \end{pmatrix}, \quad \mathcal{D}_s \in \mathbb{R}^{s\times s},$$

where $\mathcal{D}_s$ is nonsingular. Put

$$V_2 = \begin{pmatrix} V_{12} \\ V_{22} \end{pmatrix}, \quad V_{22} \in \mathbb{R}^{(p-s)\times(p-s)}.$$

From the identity $\mathcal{D}V_2 = 0$, we then get $V_{12} = 0$; consequently, V_{22} is an orthogonal matrix, and it follows that

$$V = \begin{pmatrix} V_{11} & 0 \\ 0 & V_{22} \end{pmatrix}.$$

From the CS decomposition, we then have

$$u_j = \begin{cases} \sum_{i=1}^{s} v_{ji} q_i, & j = 1, 2, \ldots, s, \\ \sum_{i=s+1}^{p} c_i v_{ji} q_i, & j = s+1, \ldots, p. \end{cases}$$

It follows that q_p is orthogonal to u_j for $j = 1, 2, \ldots, s$, and as in the cases above we can now replace u_p by q_p and increase the value of the objective function.

Thus, we have shown that for $C \neq I_p$ the stationary point does not correspond to the global maximum, which proves the lemma. □

References

Absil, P.-A., Mahony, R., & Sepulchre, R. (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton: Princeton University Press.

Adachi, K., & Trendafilov, N. (2017). Sparsest factor analysis for clustering variables: A matrix decomposition approach. *Advances in Data Analysis and Classification*, *12*, 778–794.

Aravkin, A., Becker, S., Cevher, V., & Olsen, P. (2014). A variational approach to stable principal component pursuit. In *Conference on uncertainty in artificial intelligence (UAI)*.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, *30*, 41–47.

Cai, J.-F., Candès, E. J., & Shen, Z. (2008). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, *20*, 1956–1982.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2009). Robust principal component analysis? *Journal of ACM*, *58*, 1–37.

De Leeuw, J. (2004). Least squares optimal scaling of partially observed linear systems. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 121–134). Dordrecht, NL: Kluwer Academic Publishers.

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, *20*, 303–353.

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. Philadelphia: SIAM.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations* (4th ed.). Baltimore, MD: Johns Hopkins University Press.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: University of Chicago Press.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, *12*, 531–547.

Journée, M., Nesterov, Y., Richtárik, P., & Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, *11*, 517–553.

Lin, Z., Chen, M., Wu, L., & Ma, Y. (2009). *The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices*. UIUC Technical Report, UILU-ENG-09-2215, November.

Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M., & Ma, Y. (2009). *Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix*. UIUC Technical Report, UILU-ENG-09-2214, August.

Mulaik, S. A. (2005). Looking back on the factor indeterminacy controversies in factor analysis. In In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics* (pp. 174–206). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Mulaik, S. A. (2010). *The foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low-rank matrix approximation. *Journal of Multivariate Analysis*, *99*, 1015–1034.

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels. *Psychometrika*, *44*, 157–166.

Steiger, J. H., & Schonemann, P. H. (1978). *A history of factor indeterminacy* (pp. 136–178). Chicago, IL: University of Chicago Press.

Trendafilov, N., Fontanella, S., & Adachi, K. (2017). Sparse exploratory factor analysis. *Psychometrika*, *82*, 778–794.

Trendafilov, N. T., & Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, *20*, 874–891.

Unkel, S., & Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review*, *78*, 363–382.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation. *Biostatistics*, *10*, 515–534.

Yuan, X., & Yang, J. (2013). Sparse and low-rank matrix decomposition via alternating direction methods. *Pacific Journal of Optimization*, *9*, 167–180.

Manuscript Received: 18 AUG 2017