ROBUST AND GLOBALLY SPARSE PCA VIA MAJORIZATION-MINIMIZATION AND VARIABLE SPLITTING

Hugo Brehier, Arnaud Breloy, Mohammed Nabil El Korso, Sandeep Kumar

ABSTRACT

This paper addresses the problem of robust and sparse PCA. We consider a formulation combining a M-estimation type robust subspace recovery term and a mixed norm that promotes structured sparsity in the basis vectors, which is especially interesting for joint dimension reduction and variable selection. To solve it, we propose to leverage variable splitting methods, with the crucial step then lying on the Stiefel manifold. The resolution of this subproblem, involving the orthonormality constraint, is achieved through a tailored majorization-minimization (MM) step. Numerical experiments on gene expression measurements illustrate the interest of the proposal.

***Index Terms*—** Sparse PCA, Robust Subspace Recovery, Stiefel Manifold, ADMM, Majorization-Minimization.

1. INTRODUCTION

Principal components analysis (PCA) [1] is probably the most celebrated solution to the problem of linear subspace recovery. From a demeaned data matrix consisting of n samples of dimension p, denoted $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$, PCA consists in computing $\tilde{\mathbf{X}} = \mathbf{U}_{\text{PCA}}^T \mathbf{X} \in \mathbb{R}^{k\times n}$ where $\mathbf{U}_{\text{PCA}} \in \text{St}(p,k)$[1] contains the k leftmost singular vectors of $\mathbf{X}$. This projection can serve for dimension reduction (from p to $k < p$) and exploratory analysis. In comparison with non-linear dimension reduction methods (kernel methods [2], manifold embeddings [3] and autoencoders [4]), PCA has an advantage in the interpretability that comes from the recovered basis.

The standard PCA appears as the solution of various problem, such as orthogonal regression (least-squares subspace recovery), variance maximisation or maximum likelihood [5]. Nevertheless, this solution is prone to several issues, which motivated the derivation of alternative PCA algorithms with some desired properties:

$a)$ *Robustness*: The standard PCA is sensitive to outliers, which can lead to irrelevant solutions when the sample set is partially corrupted. Numerous approaches exist to overcome this issue, such as the use of robust estimators of the covariance matrix, or robust geometric costs [6]. These methods are broadly referred to as *robust subspace recovery* [7].

$b)$ *Sparsity*: The standard PCA produces new variables that are linear combinations of all the original ones. Since these variables generally have a fundamental or physical meaning, such as assets in finance or genes in biology, a basis with only a few non-zero entries could facilitate the statistical interpretation and the variable selection. The problem of finding such vectors is referred to as *sparse PCA* and motivated numerous works (see e.g. [8–15] and references therein).

To enjoy the best of both worlds, [16] considered a robust sparse PCA formulation combining an M-estimation type fitting term [17, 18] and sparsity promoting penalties from [19]. In this paper, we propose a new algorithm suited to this approach that leverages variable splitting methods, see e.g. SOC [20] or the Manifold Alternating Direction Method of Multipliers (MADMM) [21]. Indeed, jointly handling the orthonormality constraint and sparsity promoting penalties is generally a hard task. The algorithms in [16] resort to smoothed penalties from [19] that allow to deal with the issue within the majorization-minimization framework. The proposed approach will allow for a direct use of non-smooth penalties. A notable interest is that it will newly permit the use of a mixed norm [22]. This type of penalty promotes structured sparsity patterns (e.g. *globally* sparse PCA [23, 24] or group Lasso [25]), which is useful in variable selection.

Though the proposed algorithm can be generalized to various robust costs functions (cf. [16]), this paper will focus on a Huber-type one, whose limit case yields the median subspace. The resulting algorithm will thus be referred to as median sparse PCA (MSPCA). Then, validations experiments on gene expression data will show that the proposed approach achieves a sparsity versus explained variance trade-off that is comparable to the state of the art, but without sacrificing the orthogonality of the basis vectors. It will also illustrate the interest of combining both robust subspace recovery fitting costs and sparse penalties.

In the following sections, let $\mathbf{A}$ be a matrix with $(i,j)^{th}$ entry $[\mathbf{A}]_{i,j}$, i^{th} column $\mathbf{a}_i$ and i^{th} row $\mathbf{A}_{i,:}$. Further on, we denote by $\|\cdot\|_p$ the ℓ_p norm, by $\|\cdot\|_F$ the Frobenius one and the $\ell_{2,1}$-norm by $\|\mathbf{A}\|_{2,1} \triangleq \sum_{i=1}^p \|\mathbf{A}_{i,:}\|_F$

Hugo Brehier, Arnaud Breloy are with LEME (EA4416), University Paris Nanterre, France. Mohammed Nabil El Korso is with L2S (UMR 8506), CNRS, CentraleSupélec, Université Paris-Saclay. Sandeep Kumar is with the Department of Electrical Engineering, The Indian Institute of Technology Delhi, India. This work was supported by ComUE Université Paris Lumières (UPL) and ANR MASSILIA (ANR-21-CE23-0038-01).

[1] Stiefel manifold, denoted as $\text{St}(p,k) = \{\mathbf{U} \in \mathbb{R}^{p\times k} : \mathbf{U}^T\mathbf{U} = \mathbf{I}_k\}$

2. PROBLEM FORMULATION

We consider the following robust sparse PCA problem:

$$\begin{array}{ll} \min\limits_{\mathbf{U}} & \frac{1}{n} F_{q,\delta}(\mathbf{U}, \mathbf{X}) + \lambda \psi(\mathbf{U}) \\ \text{s.t.} & \mathbf{U} \in \mathrm{St}(p, k) \end{array} \tag{1}$$

where:

- $F_{q,\delta}$ is a robust Huber-type data-fitting cost [16, 17]:

$$F_{q,\delta}(\mathbf{U}, \mathbf{X}) = \sum_{i=1}^{n} \rho(\mathrm{dist}(\mathbf{x}_i, \mathbf{U})) \tag{2}$$

with $\mathrm{dist}(\mathbf{x}_i, \mathbf{U}) = \left\|\mathbf{x}_i - \mathbf{U}\mathbf{U}^T \mathbf{x}_i\right\|_F$ and the function:

$$\rho(x) = \begin{cases} \frac{1}{2\delta} x^2 + (q\delta)^{q/(2-q)} - \frac{(q\delta)^{2/(2-q)}}{2\delta} & \text{if } x^{2-q} < q\delta \\ x^q & \text{if } x^{2-q} \geq q\delta \end{cases} \tag{3}$$

induces some robustness in the estimation process thanks to the use of a q-norm rather than a quadratic term for large errors. Typically, $q = 1$ and a small δ yield a tractable approximation of the median subspace. Otherwise, the classical least-squares subspace fitting is obtained for $q = 2$. The interest of $F_{q,\delta}$ was evidenced in robust subspace recovery but not fully leveraged in (globally) sparse PCA with standard sparsity promoting penalties such as ℓ_1 and ℓ_{21}-norms.

- $\psi(\cdot)$ is a sparsity promoting penalty and $\lambda \in \mathbb{R}^+$ is a regularization parameter. For example, the ℓ_1-norm promotes unstructured sparsity in the basis $\mathbf{U}$, while the $\ell_{2,1}$-norm promotes structured sparsity [25]. Its row-wise structure enforces the same sparsity pattern along all basis vectors, so as to reveal a subset of useful variables.

Even for the usual sparse PCA ($q = 2$ and ℓ_1-penalty), most works relax the orthogonality constraint on $\mathbf{U}$ in (1). This approach tends to create correlated loading vectors at high regularization and thus degrade the explainability of the PCs. We will leverage MADMM to avoid such relaxation.

3. MADMM ALGORITHM

The problem (1) is difficult to handle due to the orthonormality constraint on $\mathbf{U}$ and the non-smooth regularization penalty. We therefore consider a variable-splitting reformulation which will allow us to derive a tractable and efficient algorithm through MADMM [21], a special case of the renowned ADMM [26–28]. Indeed, we first formulate:

$$\begin{array}{ll} \min\limits_{\mathbf{U},\mathbf{V}} & \frac{1}{n} F_{q,\delta}(\mathbf{U}, \mathbf{X}) + \lambda \psi(\mathbf{V}) \\ \text{s.t.} & \mathbf{U} \in \mathrm{St}(p, k), \mathbf{U} = \mathbf{V}. \end{array} \tag{4}$$

We keep the constraint $\mathbf{U} \in \mathrm{St}(p, k)$ while relaxing it on $\mathbf{V}$ and conversely for the sparsity constraint. The constraint $\mathbf{V} = \mathbf{U}$ implies that $\mathbf{V}$ is eventually orthonormal and $\mathbf{U}$ sparse, which allows for deriving practical updates of both variables. The augmented Lagrangian for (4) is:

$$L(\mathbf{U}, \mathbf{V}, \mathbf{\Gamma}) = \frac{1}{n} F_{q,\delta}(\mathbf{U}, \mathbf{X}) + \lambda \psi(\mathbf{V}) + \gamma \|\mathbf{U} - \mathbf{V}\|_F^2 + \langle \mathbf{\Gamma}, \mathbf{U} - \mathbf{V} \rangle \tag{5}$$

with $\gamma \in \mathbb{R}^+$, and where $\mathbf{\Gamma} \in \mathbb{R}^{p\times k}$ contains Lagrange multipliers associated with the constraint $\mathbf{U} = \mathbf{V}$. We then aim at alternatively minimizing this objective L for each variable, while keeping other fixed. Each step is detailed below and the full algorithm is summarized in Algorithm 1.

Algorithm 1 MADMM for MSPCA

```
Entry: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$
Initialize $\mathbf{U} \in \mathbb{R}^{p\times k}$ (by PCA or another method)
Initialize $\mathbf{V} \in \mathbb{R}^{p\times k}$ randomly
Initialize $\mathbf{\Gamma} \in \mathbb{R}^{p\times k}$ by a matrix of zeros
repeat
    Update $\mathbf{U}$ by following Algorithm 2
    Update $\mathbf{V}$ by following (15) or (16)
    Update $\mathbf{\Gamma}$ by : $\mathbf{\Gamma} + 2\gamma(\mathbf{U} - \mathbf{V})$
until convergence
Output: $\mathbf{U} \in \mathbb{R}^{p\times k}$
```

3.1. U-update

This step consists in fixing $\mathbf{V}$ and $\mathbf{\Gamma}$ and solving for a variable constrained to the Stiefel manifold:

$$\mathbf{U} = \underset{\mathbf{U}\in\mathrm{St}(p,k)}{\operatorname{argmin}} \quad L(\mathbf{U}, \mathbf{V}, \mathbf{\Gamma}), \tag{6}$$

Leaving out constant terms, we then consider:

$$\begin{array}{ll} \min\limits_{\mathbf{U}} & \frac{1}{n} F_{q,\delta}(\mathbf{U}, \mathbf{X}) + \gamma \|\mathbf{U} - \mathbf{V}\|_F^2 + \langle \mathbf{\Gamma}, \mathbf{U} - \mathbf{V} \rangle \\ \text{s.t.} & \mathbf{U} \in \mathrm{St}(p, k) \end{array} \tag{7}$$

The solution to this problem admits no closed-form solution. We therefore tailor a majorization-minimization algorithm in order to evaluate a local minimum of the objective function. The majorization step is applied thanks to the following proposition:

Proposition 3.1. *The objective function in* (7) *is majorized at point* $\mathbf{U}_t$ *by the surrogate function:*

$$g(\mathbf{U}|\mathbf{U}_t) = -\operatorname{Tr}\left(\mathbf{U}^T \left(\frac{q}{n} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \mathbf{U}_t + 2\gamma \mathbf{V} - \mathbf{\Gamma}\right)\right) \tag{8}$$

with $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ \ldots, \ \tilde{\mathbf{x}}_n]$ *and* $\forall i = 1, \ldots, n$ *:*

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i / \max\left(\mathrm{dist}^{(2-q)/2}(\mathbf{x}_i, \mathbf{U}_t), \sqrt{q\delta}\right). \tag{9}$$

Equality is achieved at $\mathbf{U}_t$.

Proof. First, we leverage the result of [17], stating that $F_{q,\delta}$ can be majorized at point $\mathbf{U}_t$ by the quadratic surrogate function

$$H_{q,\delta}(\mathbf{U}, \mathbf{X}|\mathbf{U}_t) = \frac{q}{2} \sum_{i=1}^{n} \mathrm{dist}^2(\tilde{\mathbf{x}}_i, \mathbf{U}) + \mathrm{const}. \tag{10}$$

Algorithm 2 $\mathbf{U}$-Update

```
Entry: $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$
repeat
    Form $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n] \in \mathbb{R}^{p \times n}$
    with $\tilde{\mathbf{x}}_i = \mathbf{x}_i / \max\left(\mathrm{dist}^{(2-q)/2}(\mathbf{x}_i, \mathbf{U}), \sqrt{q\delta}\right)$
    Compute $\mathbf{C} = \frac{q}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{U} + 2\gamma\mathbf{V} - \mathbf{\Gamma}$
    Compute the thin SVD: $\mathbf{C} = \mathbf{U}_C \mathbf{\Sigma}_C \mathbf{V}_C^T$
    Update $\mathbf{U}$ by : $\mathbf{U} = \mathbf{U}_C \mathbf{V}_C^T$
until convergence
Output: $\mathbf{U} \in \mathbb{R}^{p \times k}$
```

with $\tilde{\mathbf{x}}_i$ defined in (9) (implicitly dependent on q and δ) and $\mathrm{const.}$ denoting a constant term w.r.t. $\mathbf{U}$. Now, remark that:

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{dist}^2(\tilde{\mathbf{x}}_i, \mathbf{U}) &= \sum_{i=1}^{n} \left\|\tilde{\mathbf{x}}_i - \mathbf{U}\mathbf{U}^T\tilde{\mathbf{x}}_i\right\|^2 \\ &= -\operatorname{Tr}(\mathbf{U}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{U}) + \mathrm{const.}\end{aligned} \tag{11}$$

Denote $f_B(\mathbf{U}) = \operatorname{Tr}(\mathbf{U}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{U})$. The function $-f_B(\mathbf{U})$ is quadratic concave, so it can be majorized at point $\mathbf{U}_t$ by its first order Taylor expansion [29]:

$$\begin{aligned}-f_B(\mathbf{U}) &\leq -f_B(\mathbf{U}_t) - \operatorname{Tr}\left(\left(\frac{\partial f_B(\mathbf{U}_t)}{\partial \mathbf{U}_t}\right)^T (\mathbf{U} - \mathbf{U}_t)\right) \\ &= -2\operatorname{Tr}(\mathbf{U}^T\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{U}_t) + \mathrm{const.}\end{aligned} \tag{12}$$

Adding the last two terms of (7), this linear majorizer yields the surrogate g as in (8). □

From the majorizer in Proposition 3.1, the minimization of the surrogate g is equivalent to:

$$\mathbf{U}_{t+1} = \underset{\mathbf{U} \in \mathrm{St}(p,k)}{\operatorname{argmax}} \operatorname{Tr}\left(\mathbf{U}^T\left(\frac{q}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{U}_t + 2\gamma\mathbf{V} - \mathbf{\Gamma}\right)\right) \tag{13}$$

which corresponds to an orthogonal Procrustes problem [30] whose solution can be computed as $\mathbf{U}_{t+1} = \mathbf{U}_C\mathbf{V}_C^T$, with the thin-SVD: $\mathbf{C} \overset{\mathrm{TSVD}}{=} \mathbf{U}_C\mathbf{\Sigma}_C\mathbf{V}_C^T$ and $\mathbf{C} = \frac{q}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{U}^t + 2\gamma\mathbf{V} - \mathbf{\Gamma}$. The resulting algorithm to compute the $\mathbf{U}$-update is summarized in the box Algorithm 2.

3.2. V-update

Thanks to the variable splitting, this step consists simply in a proximal evaluation. Fixing $\mathbf{U}$ and $\mathbf{\Gamma}$, we get:

$$\mathbf{V} = \underset{\mathbf{V}}{\operatorname{argmin}} \quad \gamma\left\|\tilde{\mathbf{U}} - \mathbf{V}\right\|_F^2 + \lambda\psi(\mathbf{V}), \tag{14}$$

with $\tilde{\mathbf{U}} = \mathbf{U} + \frac{1}{2\gamma}\mathbf{\Gamma}$. This gives the problem the form of a proximal [31] of the function ψ. If ψ is the ℓ_1-norm, the solution is given by:

$$\mathbf{V} = S_{\lambda/2\gamma}(\tilde{\mathbf{U}}) \tag{15}$$

with the entry-wise soft-thresholding operator, defined (entry by entry) by: $[S_{\lambda/2\gamma}(\tilde{\mathbf{U}})]_{ij} = \mathrm{sgn}([\tilde{\mathbf{U}}]_{ij})\left(|[\tilde{\mathbf{U}}]_{ij}| - \frac{\lambda}{2\gamma}\right)_+$ where $(x)_+$ is the positive part function and sgn is the sign function. If ψ is the $\ell_{2,1}$-norm, the solution is given by:

$$\mathbf{V} = T_{\lambda/2\gamma}(\tilde{\mathbf{U}}) \tag{16}$$

with the row-wise thresholding operator, defined (row by row) by: $[T_{\lambda/2\gamma}(\tilde{\mathbf{U}})]_{i:} = \left(1 - \frac{\lambda/2\gamma}{\|[\tilde{\mathbf{U}}]_{i:}\|}\right)_+ [\tilde{\mathbf{U}}]_{i:}$

For completeness, we mention the $\mathbf{\Gamma}$-update, a generic dual ascent step: $\mathbf{\Gamma} \leftarrow \mathbf{\Gamma} + 2\gamma(\mathbf{U} - \mathbf{V})$

4. EXPERIMENTS

4.1. Simulation study for robustness

In the following, synthetic data will be drawn according to a Gaussian probabilistic model.

Definition 4.1. *Probabilistic PCA Model (PPCA) [5]*
Samples are drawn as a structured signal plus noise :

$$\mathbf{x} = \mathbf{U}_0\mathbf{s} + \mathbf{n} \tag{17}$$

with $\mathbf{U}_0 \in \mathbb{R}^{p \times k}$, $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^k$ *and* $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \in \mathbb{R}^p$*. Then,* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *with* $\mathbf{\Sigma} = \mathbf{U}_0\mathbf{U}_0^T + \sigma^2\mathbf{I}$*.*

Next, the outliers will be generated according to a Haystack-type model, which consists in a specific mixture of PPCA models, defined as follows.

Definition 4.2. *Haystack Model [32]*
Samples $\{\mathbf{x}_i\}_{i=1}^n$ *are drawn as inliers and outliers as follows:*

$$\begin{aligned}\{\mathbf{x}_i\}_{i=1}^n &= \{\{\mathbf{x}_i^{in}\}_{i=1}^{n_{in}}, \{\mathbf{x}_i^{out}\}_{i=n_{in}+1}^n\} \\ \mathbf{x}^{in} &\sim \mathcal{N}(\mathbf{0}, \sigma_s^2\mathbf{U}_0{\mathbf{U}_0}^T + \mathbf{I}_p) \\ \mathbf{x}^{out} &\sim \mathcal{N}(\mathbf{0}, \sigma_o^2\mathbf{U}_0^{\perp}(\mathbf{U}_0^{\perp})^T + \mathbf{I}_p)\end{aligned} \tag{18}$$

where $\mathbf{U}_0$ *is the underlying signal subspace basis,* $\mathbf{U}_0^{\perp} \in \mathrm{St}(p, p-k)$ *is its orthonormal complement,* σ_s^2 *and* σ_o^2 *are respectively the signal and outlier to noise ratio while* $pc_o = (n - n_{in})/n$ *is the fraction of outliers.*

The sparse signal subspace basis $\mathbf{U}_0$ is generated as: $\mathbf{U}_0 = \begin{bmatrix}\mathbf{U}_d \\ \mathbf{0}\end{bmatrix}$, Where $\mathbf{U}_d \in \mathbb{R}^{d \times k}, d \leq p$, is an orthogonal basis and $\mathbf{0} \in \mathbb{R}^{(p-d) \times k}$ is a matrix of zeros. For some resulting basis $\mathbf{U}$, the performance is evaluated in terms of average fraction of energy:

$$\mathrm{AFE}(\mathbf{U}) = \mathbb{E}\left[\operatorname{Tr}(\mathbf{U}^T\mathbf{U}_0{\mathbf{U}_0}^T\mathbf{U})\right]/k \tag{19}$$

which assesses if the subspace spanned by $\mathbf{U}_0$ is well recovered (on average) by the estimation process. The expectation is evaluated through 250 Monte-Carlo runs.

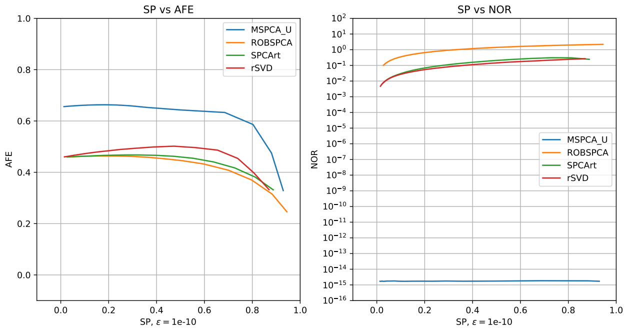

Fig. 1. AFE and NOR versus SP on synthetic data

This criterion will slightly favor the algorithms that relax the orthonormality constraint in PCA, so we also assess the non-orthogonality criterion $\mathrm{NOR}(\hat{\mathbf{U}}) = ||\hat{\mathbf{U}}^H\hat{\mathbf{U}} - \mathbf{I}||_F^2$. We measure these two criterions against the degree of sparsity of the basis: $\mathrm{SP}(\mathbf{U}) = 1 - \frac{\|\mathbf{U}\|_0}{p\times k}$. The methods we compare against are rSVD [13], a vector-by-vector method with deflation from [33], SPCArt [14], a basis-wide method, and ROBSPCA [34].

Figure 1 shows the result of such a simulation with $p = 100, k = 5, n = 100, d = 0.5p$ while $\sigma_s = \sigma_o = \sqrt{(10)}$ and $pc_o = 0.05$. MSPCA is set with $q = 1$ and $\delta = 1$ (whose value does not dramatically impacting the results, as observed in [17]). It is evident that MSPCA, thanks to its robust cost, deals with outliers better than other methods from the leftmost plot: the AFE over all sparsity degrees is higher for MSPCA. From the rightmost plot we find out that MSPCA also respects more stricly the orthonormality constraint.

4.2. Study on a microarray gene dataset

Experiments are carried out on the Khan Gene Data [35], which consists of 2308 gene expression measurements from small round blue cell tumors of 63 patients across 4 classes (cancer types). We study the sparsity-performance trade-off of MSPCA using either ℓ_1-norm or $\ell_{2,1}$-norm penalties. Since there is no ground truth for the 'true' underlying subspace, the AFE is computed as the fraction of recovered variance in $\mathbf{X}$.

Figure 2 displays the AFE and NOR versus the degree of *row*-sparsity (SProw). ROBSPCA and rSVD, not suited to structured sparsity, perform worse. We notice that MSPCA-$\ell_{2,1}$ achieves an AFE close to the one of SPCArt-$\ell_{2,1}$ (i.e., we modify SPCArt algorithm to use the $\ell_{2,1}$-norm penalty), while not relaxing the orthonormality constraints.

Finally, Figure 3 displays the first 3 principal components obtained with various algorithms: i) standard PCA; ii) MSPCA with no sparse penalty (equivalent to RSR in [17]); iii) MSPCA-ℓ_1; iv) MSPCA-$\ell_{2,1}$. All robust formulations use a Huber cost with $q = 1$ and $\delta = 1$. The sparse penalties are set to obtain a similar sparsity (SP) level.

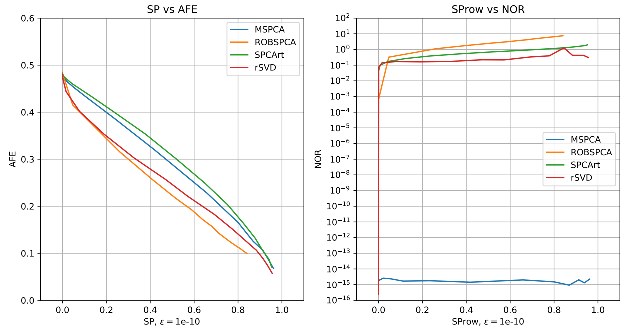

Fig. 2. AFE and NOR versus SProw of various sparse PCA algorithms on Khan Gene Data.

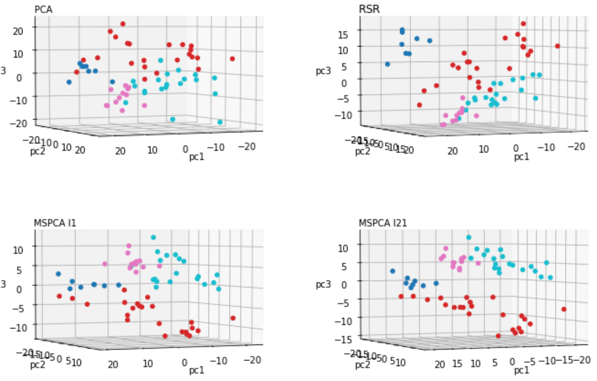

Fig. 3. Display of the 3 first principal components from PCA (top left), RSR (top right), MSPCA-ℓ_1 (bottom left), MSPCA-$\ell_{2,1}$ (bottom right).Each color in the point cloud represent a different class (cancer type).

We notice the interest of the robust fitting criterion (2) compared to the standard least squares PCA formulation, as the obtained principal components separates the classes more clearly across the data points cloud (top row). We also observe that the sparse PCA approaches offer a projection where the classes appear more clearly separated (bottom-left). The proposed $\ell_{2,1}$ penalty allows us to perform an interesting joint dimension reduction and variable selection, that improves the class separation (bottom-right) in an unsupervised manner.

5. CONCLUSION

This paper proposed a method to perform sparse PCA using a robust subspace recovery fitting and a non-smooth sparsity promoting penalty. Experiments on gene expression data showed both the interest of the approach in terms of explained variance versus sparsity trade-off, as well as for the visual representation of the principal components.

6. REFERENCES

[1] I.T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics. Springer, 2002.

[2] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller, *Kernel Principal Component Analysis*, p. 327–352, MIT Press, Cambridge, MA, USA, 1999.

[3] Sam T. Roweis and Lawrence K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[4] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[5] M. E. Tipping and C. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 21, no. 3, pp. 611–622, January 1999.

[6] P.J. Huber, *Robust Statistics*, Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 2004.

[7] G. Lerman and T. Maunu, "An overview of robust subspace recovery," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1380–1410, 2018.

[8] Ian T. Jolliffe, Nickolay T. Trendafilov, and Mudassir Uddin, "A modified principal component technique based on the lasso," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 531–547, 2003.

[9] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 517–553, Mar. 2010.

[10] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007.

[11] Zhaosong Lu and Yong Zhang, "An augmented lagrangian approach for sparse principal component analysis," *Mathematical Programming*, vol. 135, 07 2009.

[12] Hui Zou, Trevor Hastie, and Robert Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[13] Haipeng Shen and Jianhua Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1015 – 1034, 2008.

[14] Z. Hu, G. Pan, Y. Wang, and Z. Wu, "Sparse principal component analysis via rotation and truncation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 875–890, 2016.

[15] Y. Uematsu, Y. Fan, K. Chen, J. Lv, and W. Lin, "SOFAR: Large-scale association network learning," *IEEE Transactions on Information Theory*, vol. 65, no. 8, pp. 4924–4939, 2019.

[16] A. Breloy, S. Kumar, Y. Sun, and D.P. Palomar, "Majorization-minimization on the stiefel manifold with application to robust sparse PCA," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1507–1520, 2021.

[17] G. Lerman and T. Maunu, "Fast, robust and non-convex subspace recovery," *Information and Inference: A Journal of the IMA*, vol. 7, no. 2, pp. 277–336, 12 2017.

[18] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational invariant l1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, p. 281–288.

[19] K. Benidis, Y. Sun, P. Babu, and D.P Palomar, "Orthogonal sparse pca and covariance estimation via procrustes reformulation," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6211–6226, 2016.

[20] R. Lai and S. Osher, "A splitting method for orthogonality constrained problems," *J. Sci. Comput.*, vol. 58, no. 2, pp. 431–449, Feb. 2014.

[21] A. Kovnatsky, K. Glashoff, and M.M. Bronstein, "MADMM: a generic algorithm for non-smooth optimization on manifolds," in *European Conference on Computer Vision*. Springer, 2016, pp. 680–696.

[22] Matthieu Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303 – 324, 2009.

[23] Pierre-Alexandre Mattei, Charles Bouveyron, and Pierre Latouche, "Globally sparse probabilistic pca," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 976–984.

[24] Abd-Krim Seghouane, Navid Shokouhi, and Inge Koch, "Sparse principal component analysis with preserved sparsity pattern," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3274–3285, 2019.

[25] Ming Yuan and Yi Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, pp. 49–67, 2006.

[26] P. L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[27] Jonathan Eckstein and Dimitri Bertsekas, "On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 04 1992.

[28] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[29] Y. Sun, P. Babu, and D. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. on Signal Process.*, vol. PP, no. 99, pp. 1–1, 2016.

[30] Gene H Golub and Charles F Van Loan, *Matrix computations*, vol. 3, JHU press, 2012.

[31] N. Parikh, S. Boyd, et al., "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[32] Gilad Lerman, Michael B. McCoy, Joel A. Tropp, and Teng Zhang, "Robust computation of linear models, or how to find a needle in a haystack," *CoRR*, vol. abs/1202.4044, 2012.

[33] Lester W. Mackey, "Deflation methods for sparse pca," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., pp. 1017–1024. Curran Associates, Inc., 2009.

[34] N. Benjamin Erichson, Peng Zheng, Krithika Manohar, Steven L. Brunton, J. Nathan Kutz, and Aleksandr Y. Aravkin, "Sparse principal component analysis via variable projection," *SIAM Journal on Applied Mathematics*, vol. 80, no. 2, pp. 977–1002, 2020.

[35] J. Khan, J.S. Wei, M. Ringner, L.H Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[36] Davood Hajinezhad and Mingyi Hong, "Nonconvex alternating direction method of multipliers for distributed sparse principal component analysis," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015, pp. 255–259.