Sparse Principal Component Analysis via Variable Projection

N. Benjamin Erichson[*], Peng Zheng[†], Krithika Manohar[‡], Steven L. Brunton[‡], J. Nathan Kutz[‡], and Aleksandr Y. Aravkin[‡]

Abstract. Sparse principal component analysis (SPCA) has emerged as a powerful technique for data analysis, providing improved interpretation of low-rank structures by identifying localized spatial structures in the data and disambiguating between distinct time scales. We demonstrate a robust and scalable SPCA algorithm by formulating it as a value-function optimization problem. This viewpoint leads to a flexible and computationally efficient algorithm. It can further leverage randomized methods from linear algebra to extend the approach to the large-scale (big data) setting. Our proposed innovation also allows for a robust SPCA formulation which can obtain meaningful sparse components in spite of grossly corrupted input data. The proposed algorithms are demonstrated using both synthetic and real world data, showing exceptional computational efficiency and diagnostic performance.

Key words. Dimension reduction, sparse principal component analysis, randomized algorithms.

AMS subject classifications. 62H25, 62J12, 68W20

1. Introduction. A wide range of phenomena in the physical, engineering, biological, and social sciences feature rich dynamics that give rise to multiscale structures in both space and time, including fluid dynamics, atmospheric-ocean interactions, climate modeling, epidemiology, and neuroscience. Remarkably, the underlying dynamics of such systems are typically inherently low-rank in nature, generating data sets where dimensionality reduction techniques, such as principal component analysis (PCA), can be used as a critically enabling diagnostic tool for interpretable characterizations of the dynamics. PCA decompositions express time-varying patterns as a linear combination of the dominant correlated spatial activity of the state of a system as it evolves in time. Although commonly used, the PCA approach generates global modes that often mix or blend various spatio-temporal scales, and cannot identify underlying governing dynamics that act at separate scales. Moreover, classic PCA also tends to overfit data where the number of observations is smaller than the number of variables [11].

Constrained or regularized matrix decompositions provide a more flexible approach for modeling dynamic patterns. Specifically, prior information can be introduced through sparsity promoting regularizers to obtain a more parsimonious approximation of the data which typically provides improved interpretability. Among others, *sparse principal component analysis* (SPCA) has emerged as a popular and powerful technique for modern data analysis. SPCA promotes sparsity in the modes, i.e., the sparse modes have only a few *active* coefficients, while the majority of coefficients are constrained to be zero. The resulting sparse modes are often highly localized and more interpretable than the global PCA modes obtained from traditional PCA. As a consequence, sparse regularization of PCA allows for a decomposition strategy that can specifically identify localized spatial structures in the data and disambiguate between distinct

[*]Department of Statistics, University of California Berkeley, Berkeley, CA. (Email: erichson@berkeley.edu)
[†]Department of Applied Mathematics, University of Washington, Seattle, WA.
[‡]Department of Mechanical Engineering, University of Washington, Seattle, WA.

time scales, both of which are ubiquitous in measurement data of complex systems. As an example, one only needs to consider the physical phenomenon of El Nino which is a mode characterized by a localized warm temperature profile which traverses the southern Pacific ocean. This is a highly localized mode that, as will be shown, is well characterized by SPCA, while standard PCA gives a global mode with nonzero values across the entire globe.

While the idea of sparsifying the weight vectors is not new, simple ad-hoc techniques such as naive thresholding can lead to misleading results. A formal approach to SPCA, using ℓ_1 regularization, was first proposed by Jolliffe et al. [31]. This pioneering work lead to a variety of sparsity promoting algorithms [59, 17, 19, 49, 48, 56, 32]. The success of sparse PCA in obtaining interpretable modes motivates the general approach developed in this paper. Specifically, our method offers three immediate improvements over previously proposed SPCA algorithms: (1) a faster and scalable algorithm, (2) robustness to outliers, and (3) straightforward extension to nonconvex regularizers, including the ℓ_0 norm. Scalability is essential for many applications — for example, dynamical systems generate very large-scale datasets, such as the sea surface temperature data analyzed in this paper. Robust formulations allow SPCA to be deployed in a broader setting, where data contamination could otherwise hide sparse modes. Nonconvex regularizers are not currently available in SPCA software — we show that the modes we get with these approaches are better in synthetic examples, and more interpretable for real data.

Contributions of this work. In this work, we develop a scalable and robust approach for SPCA. A key feature of the approach is the use of *variable projection* to partially minimize over the orthogonally constrained variables. This idea was used in the original alternating approach of [59], and we innovate on this idea by recasting the problem as a value-function optimization. This viewpoint allows for significantly faster algorithms, scalability, and broader applicability. We also allow nonconvex regularization on the loadings, which further improves interpretability and sparsity. Not only does the method scale well, but it is further accelerated using randomized methods for linear algebra [22]. Further, the proposed approach extends to robust SPCA formulations, which can obtain meaningful principal components even with grossly corrupted input data. The outliers are modeled as perturbations to the data, as in the robust PCA model [12, 7, 1]. These innovations provide a flexible and highly-efficient algorithm for modern data analysis and diagnostics that will enable a wide range of critical applications at a scale not previously possible with other leading algorithms.

Organization. The manuscript is organized as follows: Section 2 reviews PCA and the variable projection framework. Section 3 provides a detailed problem formulation and discusses the variable projection viewpoint which is advocated in this paper. Further, different loss functions and regularizes are discussed. We present the details of the proposed algorithms in Section 4. First, the standard case, using the least squares loss function, is discussed. Then, a randomized accelerated and a robust algorithm is presented. The method is applied to several examples in Section 6, where SPCA correctly identifies dynamics occurring at different timescales in multiscale data. We draw conclusions about the method and discuss its outlook in Section 7.

Notation. Scalars are denoted by lower case letters x, and vectors in $\mathbb{R}^n$ are denoted as bold lower case letters $\mathbf{x} = [x_1, x_2, ..., x_n]^\top$. Matrices are denoted by bold capital letters $\mathbf{M}$. The transpose of a real matrix is denoted as $\mathbf{M}^\top$. The spectral or operator norm of a matrix is denoted as $\| \cdot \|$ and the Frobenius norm is denoted as $\| \cdot \|_F$.

2. Background.

2.1. Principal Component Analysis.

Principal component analysis (PCA) is an ubiquitous dimension reduction technique, tracing back to Pearson [44] and Hotelling [27]. The aim of PCA is to find a set of new uncorrelated variables, called principal components (PCs), such that the first PC accounts for the greatest amount of variance in the data, the second PC for the second greatest variance, and so on. More concretely, let $\mathbf{X}$ be a real data matrix of dimension $n \times p$, with column-wise zero empirical mean. The n rows represent observations and the p columns correspond to measurements of variables. The principal components $\mathbf{z}_i \in \mathbb{R}^n$ are formed as a linear weighted combination of the variables

$$(2.1) \qquad \mathbf{z}_i = \mathbf{X}\mathbf{a}_i,$$

where $\mathbf{a}_i \in \mathbb{R}^p$ is a vector of weights. This can be expressed more concisely as

$$(2.2) \qquad \mathbf{Z} = \mathbf{X}\mathbf{A},$$

with $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_p] \in \mathbb{R}^{n \times p}$ and $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p] \in \mathbb{R}^{p \times p}$. The orthonormal matrix $\mathbf{W}$ rotates the data into a new space, where the principal components sequentially capture the maximum variability in the input data. The columns of $\mathbf{A}$ are also often denoted as modes, basis functions, principal direction or loadings.

Mathematically, a variance maximization problem can be formulated to find the weight vectors $\mathbf{a}_i$. Alternatively, the problem can be formulated as a least-squares problem, i.e., minimizing the sum of squared residual errors between the input and the projected data

$$(2.3) \qquad \begin{aligned} \underset{\mathbf{A}}{\text{minimize}} \quad & f(\mathbf{A}) = \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}^\top\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \mathbf{A}^\top\mathbf{A} = \mathbf{I}, \end{aligned}$$

where PCA imposes orthogonality constraints on the weight matrix $\mathbf{A}$. Given the singular value decomposition (SVD) of the centered (standardized) input matrix $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top,$$

the minimizer of (2.3) is given by the right singular vectors $\mathbf{V}$, i.e., we can set $\mathbf{A} = \mathbf{V}$. Further, the principal components are the scaled left singular vectors $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}$, where the entries of the diagonal matrix $\boldsymbol{\Sigma}$ are the singular values. In most applications, we are only interested in the first k dominant PCs which account for most of the variability in the input data. Thus, PCA allows one to reduce the dimensionality from p to k by simply truncating the SVD. The dominant k PCs can be used to visualize the data in low-dimensional space, and as features for clustering, classification and regression.

We refer the reader to [30] for an extensive treatment of PCA and its mechanics. Many extensions such as Kernel PCA have been proposed to extend and overcome some of the shortcomings of PCA, see [16] for an brief overview.

2.2. Variable Projection. Consider any objective of the form

$$(2.4) \qquad\qquad \min_{\mathbf{A},\mathbf{B}} g(\mathbf{A},\mathbf{B}).$$

A classic example is the nonlinear least squares problem

$$(2.5) \qquad\qquad \min_{\mathbf{A},\mathbf{B}} \quad \frac{1}{2}\|\mathbf{X} - \mathbf{\Phi}(\mathbf{B})\mathbf{A}\|^2.$$

The term 'variable projection' [23] originally arose from the fact that the least squares projection of $\mathbf{X}$ onto the range of $\mathbf{\Phi}(\mathbf{B})$ has a closed form solution, which is used explicitly in iterative methods to optimize for $\mathbf{B}$. More generally, the word 'projection' is now associated with epigraphical projection [46], or partial minimization. We can rewrite (2.4) as a value function optimization problem:

$$(2.6) \qquad\qquad \min_{\mathbf{B}} \left\{ v(\mathbf{B}) := \min_{\mathbf{A}} g(\mathbf{A},\mathbf{B}) \right\}.$$

In many cases, the function $v(\mathbf{B})$ has an explicit form. In the classic problem (2.5), we have

$$v(\mathbf{B}) = \frac{1}{2}\|\mathbf{X}(\mathbf{I} - \mathcal{P}_{\mathcal{R}(\mathbf{\Phi}(\mathbf{B}))})\|^2 = \mathrm{dist}^2(\mathbf{X}|\mathcal{R}(\mathbf{\Phi}(\mathbf{B}))),$$

where $\mathcal{P}_{\mathcal{R}(\mathbf{\Phi}(\mathbf{B}))}$ is a projector on the range of $\mathbf{\Phi}(\mathbf{B})$. Explicit expressions are not necessary as long as we have an efficient routine to compute

$$\mathbf{A}(\mathbf{B}) = \arg\min_{\mathbf{A}} g(\mathbf{A},\mathbf{B}).$$

For many problems, we can find first and second derivatives of $v(\mathbf{B})$. For example, when g is smooth and $\mathbf{A}(\mathbf{B})$ is unique, we have

$$\nabla v(\mathbf{B}) = \partial_{\mathbf{B}} g(\cdot,\cdot)|_{(\mathbf{A}(\mathbf{B}),\mathbf{B})}.$$

Formulas for second derivatives are collected in [2]. Variable projection was recently used to solve a range of large-scale structured problems in PDE-constrained optimization, nuisance parameter estimation, exponential fitting, and optimized dynamic mode decomposition [3, 2, 26, 4].

3. Problem Formulation for Sparse Principal Component Analysis (SPCA). Sparse PCA aims to find a set of sparse weight vectors, i.e., weight vectors with only a few 'active' (nonzero) values. In this manuscript, we build up on the seminal work by Zou, Hastie and Tibshirani [59], who treat SPCA as an regularized regression-type problem. More concretely, their formulation directly incorporates sparsity inducing regularizers into the optimization problem:

$$(3.1) \qquad \begin{aligned} & \underset{\mathbf{A},\mathbf{B}}{\text{minimize}} \quad f(\mathbf{A},\mathbf{B}) = \tfrac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|_{\mathrm{F}}^2 + \psi(\mathbf{B}) \\ & \text{subject to} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I}, \end{aligned}$$

where $\mathbf{B}$ is a sparse weight matrix and $\mathbf{A}$ is an orthonormal matrix. ψ denotes a sparsity inducing regularizer such as the LASSO (ℓ_1 norm) or the elastic net (a combination of the ℓ_1 and squared ℓ_2 norm). The optimization problem is minimized using an alternating algorithm:

- **Update A.** With $\mathbf{B}$ fixed, we find an orthonormal matrix $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ which minimizes

$$\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|_F^2.$$

  This is the orthogonal Procrustes problem [24] (see Appendix A), which has a closed form solution $\mathbf{A}^* = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{X}^\top\mathbf{X}\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$.

- **Update B.** With $\mathbf{A}$ fixed, we solve the optimization problem

$$\min_{\mathbf{B}} \quad \tfrac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|_F^2 + \psi(\mathbf{B}).$$

  The problem splits across the k columns of $\mathbf{B}$, yielding a regularized regression problem in each case:

$$\mathbf{b}_j^* = \arg\min_{\mathbf{b}_j} \tfrac{1}{2}\|\mathbf{X}\mathbf{A}(:,j) - \mathbf{X}\mathbf{b}_j\|^2 + \psi(\mathbf{b}_j).$$

The principal components are then formed as a sparsely weighted linear combination of the observed variables $\mathbf{Z} = \mathbf{X}\mathbf{B}$. The data can be approximately rotated back as $\widetilde{\mathbf{X}} = \mathbf{Z}\mathbf{A}^\top$.

Coordinate descent or least angle regression (LARS) are used to solve each of the k subproblems [20]. The $\mathbf{B}$ update relies on solving a strongly convex problem, and in particular the update is unique, and the algorithm as described converges to as stationary point by the analysis of [55]. Replacing ψ with a nonconvex regularizer, such as $\psi(\mathbf{B}) = \alpha\|\mathbf{B}\|_0 + \beta\|\mathbf{B}\|^2$, makes it difficult to guarantee anything about the $\mathbf{B}$ update. However, as we show, using the value function (2.6) from the variable projection viewpoint yields an efficient implementation and a straightforward convergence analysis.

3.1. Variable Projection Viewpoint. The $\mathbf{A}$ update in the method of [59] is in closed form, while the $\mathbf{B}$ update requires an iterative method. To exploit the efficiency of the $\mathbf{A}$ update, we think of projecting out $\mathbf{A}$ and introduce the sparse PCA value function

$$v(\mathbf{B}) := \min_{\mathbf{A}} \quad \tfrac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|_F^2 \quad \text{subject to} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I},$$

viewing the original SPCA problem (3.1) as

$$(3.2) \qquad\qquad\qquad \min_{\mathbf{B}} v(\mathbf{B}) + \psi(\mathbf{B}).$$

We show that $v(\mathbf{B})$ is differentiable with a Lipschitz continuous gradient, and derive its explicit form. This viewpoint lets us use any proximal algorithm we want to minimize (3.2), including proximal gradient (see e.g. [43]) and FISTA [5], with the caveat that an $\mathbf{A}$ update is computed every time $v(\mathbf{B})$ is evaluated. For the original SPCA problem, this approach rebalances the work between the $\mathbf{A}$ and $\mathbf{B}$ updates, using a single operator to update $\mathbf{B}$ instead of an iterative routine. When combined with randomized techniques for computing the $\mathbf{A}$ update, we get an order of magnitude acceleration compared to current SPCA software.

The variable projection viewpoint (3.2) also allows a robust SPCA approach with the Huber loss function. Simply replacing the quadratic penalty in (3.1) with a different loss would destroy the efficient structure of the $\mathbf{A}$ update, requiring an iterative routine to solve for it. Instead, we

use a special characterization of the Huber function to obtain a formulation with three rather than two variables, preserving the efficiency of each update. We extend our analysis to this case, so the robust formulation can also be used with any prox-friendly ψ regularizer, including the nonconvex example discussed above.

In summary, the value function viewpoint also makes it easy to extend to a broader problem setting, and we consider the following objective:

$$(3.3) \qquad \min_{\mathbf{A,B}} \quad f(\mathbf{A,B}) := \rho(\mathbf{X} - \mathbf{XBA}^\top) + \psi(\mathbf{B}) \quad \text{subject to} \quad \mathbf{A}^\top \mathbf{A} = \mathbf{I},$$

where ρ is a separable loss, while ψ is a separable regularizer for $\mathbf{B}$.

3.2. Regularizers for Sparsity. The SPCA framework incorporates a range of sparsity-inducing regularizers ψ. Sparsity is achieved by introducing additional information into the model to find the most meaningful 'active' (non-zero) entries in $\mathbf{B}$, while most of the loadings are constrained to be zero. Sparse approaches work well when many variables are redundant, i.e., not required to capture the underlying coherent model structure. Regularization also prevents overfitting, and provides a path to solve ill-posed problems, frequently encountered in the analysis of high-dimensional datasets.

3.3. Unstructured Sparsity. The ℓ_0 'norm', denoted $\ell_0(\boldsymbol{x})$ or $\|\boldsymbol{x}\|_0$, counts the number of non-zero elements in a vector $\mathbf{x}$. When used as a regularizer ψ, it encourages models with small cardinality, i.e., a small number of active loadings. Although ℓ_0 is non-smooth and non-convex, its proximal operator is simply hard thresholding (see Table 1).

In many applications, the ℓ_1 norm is used to approximate ℓ_0. In the context of least squares problems, using ℓ_1 is known as LASSO (least absolute shrinkage and selection operator). The proximal operator of the scaled ℓ_1 norm $\gamma\|\boldsymbol{x}\|_1$ is the *soft-thresholding* operator, see Table 1.

One drawback of the ℓ_1 norm is that it tends to activate only one coefficient from any set of highly correlated variables. The elastic net, introduced by Zou and Hastie [58], overcomes this drawback, using a linear combination of the ℓ_1 and quadratic penalty:
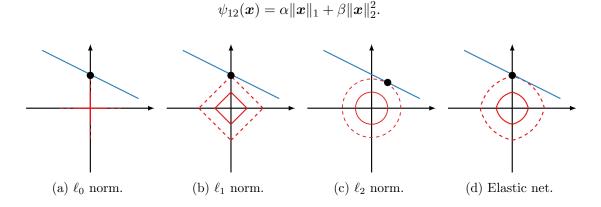
$$\psi_{12}(\boldsymbol{x}) = \alpha\|\boldsymbol{x}\|_1 + \beta\|\boldsymbol{x}\|_2^2.$$

     (a) ℓ_0 norm.   (b) ℓ_1 norm.   (c) ℓ_2 norm.   (d) Elastic net.

Figure 1: Illustration of some norms which are used as regularizers. ℓ_0, ℓ_1 and elastic net are sparsity-inducing.

	Regularizer ψ	$\mathrm{prox}_{\gamma\psi}(\boldsymbol{x})_i$	
ψ_0	$\|\boldsymbol{x}\|_0$	Hard Thresholding :	$\begin{cases} x_i, & x_i^2 > 2\gamma \\ 0, & \text{otherwise} \end{cases}$
ψ_1	$\|\boldsymbol{x}\|_1$	Soft Thresholding :	$\begin{cases} x_i - \gamma, & x_i > \gamma \\ x_i + \gamma, & x_i < -\gamma \\ 0, & \text{otherwise} \end{cases}$
ψ_{02}	$\alpha\|\boldsymbol{x}\|_0 + \beta\|\boldsymbol{x}\|_2^2$	Scaled Hard Thresholding :	$\begin{cases} x_i/(1+2\gamma\beta), & x_i^2 > 2\gamma\alpha(1+2\gamma\beta) \\ 0, & \text{otherwise} \end{cases}$
ψ_{12}	$\alpha\|\boldsymbol{x}\|_1 + \beta\|\boldsymbol{x}\|_2^2$	Scaled Soft Thresholding :	$\begin{cases} (x_i - \gamma\alpha)/(1+2\gamma\beta), & x > \gamma\alpha \\ (x_i + \gamma\alpha)/(1+2\gamma\beta), & x < -\gamma\alpha \\ 0, & \text{otherwise} \end{cases}$

Table 1: Regularizers ψ and their proximal operators.

The elastic net has an implicit grouping effect that is particularly useful for the analysis of high-dimensional multiscale physical systems, where we want to find all the associated variables which correspond to an underlying mode, rather than selecting only one variable from each underlying mode. The proximal operator of ψ_{12} combines scaling and soft thresholding, see Table 1. Following the same idea, we can also combine ℓ_0 and the quadratic penalty:

$$\psi_{02}(\boldsymbol{x}) = \alpha\|\boldsymbol{x}\|_0 + \beta\|\boldsymbol{x}\|_2^2.$$

The ψ_{02} regularizer detects correlated sets of very sparse predictors, and its proximal operator of ψ_{02} combines scaling and hard thresholding, see Table 1. Figure 1 illustrates these regularizers. Many other examples of proximal operators are collected in [13].

3.4. Structured Sparsity. A large number of separable structured regularizers ψ can be used in the proposed SPCA framework. Separability ensures that the prox-operator can be computed either in closed form or using a routine for both convex and nonconvex regularizers. Here we highlight two examples.

In some applications, selection occurs between groups of variables known *a priori*. The group lasso regularizer [57] enforces that all the variables corresponding to these predefined group are either activated or set to 0. Its prox operator can be written as

$$\mathrm{prox}_{\gamma\|\cdot\|_2}(\boldsymbol{x}) = \begin{cases} (1 - \gamma/\|\boldsymbol{x}\|_2)\boldsymbol{x}, & \|\boldsymbol{x}\|_2 > \gamma \\ \boldsymbol{0}, & \|\boldsymbol{x}\|_2 \leq \gamma \end{cases}.$$

An extension is the sparse group lasso [50], which adds an additional ℓ_1 penalty for each group. Another useful regularizer is the fused lasso [54], which gives a way to incorporate information about spatial or temporal structure in the data.

4. Fast Algorithms for Sparse PCA.

4.1. Sparse PCA via Variable Projection.

As a standard problem, we discuss the variable projection algorithm for (3.3) using the least squares loss function. We partially minimize in $\mathbf{A}$ to obtain the *value* function

$$(4.1) \qquad v(\mathbf{B}) := \min_{\mathbf{A}} \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|_F^2 \quad \text{subject to} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I}.$$

Evaluating this value function given $\mathbf{B}$ reduces to solving the orthogonal Procrustes problem [24], with closed form solution

$$(4.2) \qquad \mathbf{A}(\mathbf{B}) = \mathbf{U}\mathbf{V}^\top$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{X}^\top\mathbf{X}\mathbf{B} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, see Appendix A. Variable projection takes advantage of this closed form solution. Partially minimizing in $\mathbf{A}$ via the SVD has additional advantages over using an iterative algorithm when $\mathbf{A}$ is ill-conditioned. This is an important consideration for robust penalties, where a closed form solution for $\mathbf{A}$ is not immediately available, see Section 4.3.

The SPCA problem (3.3) is nonconvex, and so is the value function $v(\mathbf{B})$. To better understand $v(\mathbf{B})$, we consider the following simple 2D example

$$f(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{2} \|\mathbf{X} - \mathbf{X}\boldsymbol{b}\boldsymbol{a}^\top\|_F^2 \quad \text{subject to} \quad \mathbf{a}^\top\mathbf{a} = 1,$$

where $\mathbf{X} \in \mathbb{R}^{2\times2}$, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^2$. We write $v(\mathbf{b}) : \mathbb{R}^2 \to \mathbb{R}$ explicitly as

$$v(\boldsymbol{b}) = \frac{1}{2} \|\mathbf{X} - \mathbf{X}\boldsymbol{b}\,\boldsymbol{a}(\boldsymbol{b})^\top\|_F^2, \quad \boldsymbol{a}(\boldsymbol{b}) = \frac{\mathbf{X}^\top\mathbf{X}\boldsymbol{b}}{\|\mathbf{X}^\top\mathbf{X}\boldsymbol{b}\|}.$$

Figure 2 shows the level sets of this function, which are clearly nonconvex. We also see that $v(\mathbf{b})$ is smooth except at $\mathbf{b} = 0$.

We apply proximal gradient methods (see e.g. [43]) to find a stationary point of the value function $v(\mathbf{b})$ (4.1). It is easy to both evaluate $v(\mathbf{b})$ and to compute the gradient. We obtain $\mathbf{a}(\mathbf{b})$ using (4.2) and then use the formula

$$\nabla v(\mathbf{b}) = \nabla_{\mathbf{b}} f(\mathbf{a}, \mathbf{b})|_{\mathbf{a}=\mathbf{a}(\mathbf{b})} = \mathbf{X}^\top(\mathbf{X} - \mathbf{X}\boldsymbol{b}\,\boldsymbol{a}(\boldsymbol{b})^\top)\boldsymbol{a}(\boldsymbol{b}).$$
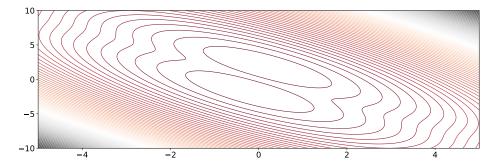

Figure 2: Level set of simple 2D projected function.

This yields a simple and efficient algorithm detailed in Algorithm 4.1. The following theorem provides a sublinear convergence guarantee for Algorithm 4.1.

Theorem 4.1 (Convergence of 4.1). *Assume $\rho = \rho_F$, then the optimality criterion satisfies*

$$\min_{1 \leq k \leq N} T(\mathbf{A}_k, \mathbf{B}_k) \leq \frac{2(\|\mathbf{X}\|_2^2 + L)^2}{N\|\mathbf{X}\|_2^2} f(\mathbf{A}_1, \mathbf{B}_1),$$

where T is the stationarity of the objective (defined in Appendix B) and L is the Lipchitz constant for $\partial\psi$.

See Appendix B.3 for the proof.

Algorithm 4.1 Variable projected proximal gradient method for (4.1)

Input: $\mathbf{A}_0$, $\mathbf{B}_0$, $k = 0$, $\epsilon > 0$, $\gamma = 1/\|\mathbf{X}\|_2^2$

  1: **while** $T(\mathbf{A}_k, \mathbf{B}_k) \geq \epsilon$ **do**   ▷ See (B.4)

  2:     $\mathbf{B}_{k+1} \leftarrow \text{prox}_{\gamma r}(\mathbf{B}_k - \gamma\mathbf{X}^\top(\mathbf{X}\mathbf{B}_k - \mathbf{X}\mathbf{A}_k))$ ▷ See (B.3)

  3:     $\mathbf{A}_{k+1} \leftarrow \arg\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}_{k+1}\mathbf{A}^\top\|_F^2$ subject to $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ ▷ See (4.2)

  4:     $k \leftarrow k + 1$

  5: **end while**

Output: $\mathbf{A}_{k+1}$, $\mathbf{B}_{k+1}$

4.2. Randomized Sparse PCA. Randomization allows efficient computation of low-rank approximations such as the SVD and PCA [25, 36, 18, 22]. We form a low-dimensional sketch (representation) of the data, which aims to capture the essential information of the original data. Using this idea, we can reformulate (3.3) as

$$(4.3) \qquad v(\mathbf{B}) := \min_{\mathbf{A}} \frac{1}{2}\|\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{B}\mathbf{A}^\top\|_F^2 \quad \text{subject to} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I}.$$

where $\widetilde{\mathbf{X}} \in \mathbb{R}^{l \times p}$ denotes the sketch of $\mathbf{X} \in \mathbb{R}^{n \times p}$. Here, the dimension l is chosen slightly larger than the target-rank k. We proceed by forming a sample matrix $\mathbf{Y} \in \mathbb{R}^{n \times l}$:

$$(4.4) \qquad \mathbf{Y} = \mathbf{X}\mathbf{\Omega},$$

where $\mathbf{\Omega} \in \mathbb{R}^{p \times l}$ is a randomly generated test matrix [25]. Next, an orthonormal basis matrix is obtained by computing the QR-decomposition of the samples matrix $\mathbf{Y} = \mathbf{Q}\mathbf{R}$. Finally, the sketch is formed by projecting the input matrix to the range of $\mathbf{Y}$, which is low-dimensional:

$$(4.5) \qquad \widetilde{\mathbf{X}} = \mathbf{Q}^\top\mathbf{X}.$$

This approach is suitable for input matrices with low-rank structure. The computational advantage becomes significant when the intrinsic rank of the data is relatively small compared to the dimension of the ambient measurement space. The quality of the sketch can be improved by computing additional power iterations [25, 22], especially if the singular value spectrum of $\mathbf{X}$ is only slowly decaying. We suggest computing at least two power iterations by default.

4.3. Robust Sparse PCA via Variable Projection. Classically, SPCA is formulated as a least-squares problem, however, it is well-known that the squared loss is sensitive to outliers. In many real world situations we face the challenge that data are grossly corrupted due to measurement errors or other effects. This motivates the need of robust methods which can more effectively account for corrupt or missing data. Indeed, several authors have proposed a robust formulation of SPCA, using the ℓ_1 norm as a robust loss function, to deal with grossly corrupted data [39, 15, 29].

For a robust formulation of SPCA, we use a closely related idea of separating a data matrix into a low-rank model and a sparse model. The architecture is depicted in Figure 3. This form of additive decomposition is well-known as robust principal component analysis (RPCA), and its remarkable ability to separate high-dimensional matrices into low-rank and sparse component makes RPCA an invaluable tool for data science [12, 7, 1]. Specifically, we suggest to use the Huber loss function $\rho = \rho_{\mathrm{H}}$ rather than the ℓ_1 norm as the data misfit. The Huber norm overcomes some of the shortcomings for the Frobenius norm and can be used as a more robust measure of fit [28, 38]. We define the Huber loss function as

$$\rho_{\mathrm{H}}(x;\kappa) = \begin{cases} \kappa|x| - \kappa^2/2, & |x| > \kappa \\ x^2/2, & |x| \le \kappa \end{cases},$$

$$\rho_{\mathrm{H}}(\mathbf{A};\kappa) = \sum_{i,j} \rho_{\mathrm{H}}(\mathbf{A}_{ij};\kappa).$$

Figure 4 illustrates the least squares and the Huber loss functions. The Huber loss functions grows at a linear rate for residuals outside the thresholding parameter κ, rather than quadrati-
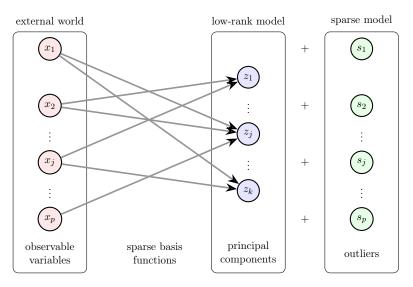

Figure 3: Robust SPCA combines a low-rank and sparse model to represent the observable variables. The low-rank model forms the principal components as a sparsely weighted linear combination of the observed variables. The sparse model captures outliers in the data.

cally. Hence, the influence of large deviations on the parameters is reduced. This is consistent with using a heavy tail distribution to model measurement errors.

The Huber penalty can be characterized as the (scaled) Moreau envelope of the ℓ_1 norm, see Section B.1.4:

$$(4.6) \qquad \rho_{\mathrm{H}}(x;\kappa) = \min_{s} \frac{1}{2}\|s - x\|^2 + \kappa\|s\|_1.$$

This characterization explicitly brings out outliers s as sparse perturbations to the data. It also makes it possible to develop efficient algorithms for the robust case. In general, our approach applies to any robust norm that can be characterized as the Moreau envelope of a separable penalty.

A naive approach loses the closed form of $\mathbf{A}$ (4.2). To preserve the advantages of partial minimization, we must place the Huber loss on the Procrustean bed of the orthogonal Procrustes problem. We use the Moreau characterization (4.6) to explicitly model sparse outliers using the variable $\mathbf{S}$, and rewrite Eq. (3.3) as follows:

$$(4.7)$$
$$\min_{\mathbf{A},\mathbf{B},\mathbf{S}} \quad f_{\mathrm{H}}(\mathbf{A},\mathbf{B},\mathbf{S}) := \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^{\top} - \mathbf{S}\|_{\mathrm{F}}^2 + \psi(\mathbf{B}) + \kappa\|\mathbf{S}\|_1 \quad \text{subject to} \quad \mathbf{A}^{\top}\mathbf{A} = \mathbf{I}.$$

Now we can again use the orthogonal Procrustes approach (4.2) and reduce (4.7) to minimizing the value function

$$(4.8) \qquad \min_{\mathbf{B},\mathbf{S}} \quad v_{\mathrm{H}}(\mathbf{B},\mathbf{S}) := \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}(\mathbf{B},\mathbf{S})^{\top} - \mathbf{S}\|_{\mathrm{F}}^2 + \psi(\mathbf{B}) + \kappa\|\mathbf{S}\|_1,$$

where $\mathbf{A}(\mathbf{B},\mathbf{S})$ is given by $\mathbf{U}\mathbf{V}^{\top}$ with

$$(4.9) \qquad (\mathbf{X} - \mathbf{S})^{\top}\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}.$$

Problem (4.8) has the same structure as (4.1) in the variables $(\mathbf{B},\mathbf{S})$, and we can easily modify the algorithm to account for the additional block, as detailed in Algorithm 4.2. The partial minimization of $\mathbf{S}$ is a prox evaluation of the 1-norm, which is the soft thresholding operator, see Table 1.

(a) Loss functions.   (b) Influence functions (first derivatives).
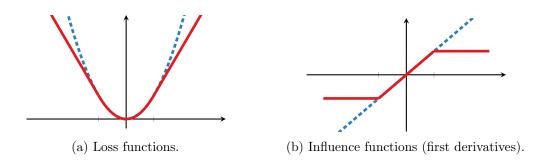
Figure 4: Illustration of the least-squares loss (dashed blue) and Huber (solid red) loss functions in (a); the first derivatives in (b) can be viewed as influence functions of the residuals.

Algorithm 4.2 Gauss-Seidel proximal gradient method for (4.7)

Input: $\mathbf{A}_0, \mathbf{B}_0, \mathbf{S}_0, k = 0, \gamma = 1/\|\mathbf{X}\|_2^2$

1: **while** not converged **do**

2: $\quad \mathbf{B}_{k+1} \leftarrow \text{prox}_{\gamma r}(\mathbf{B}_k - \gamma \mathbf{X}^\top(\mathbf{X}\mathbf{B}_k - \mathbf{X}\mathbf{A}_k + \mathbf{S}_k\mathbf{A}_k))$ $\qquad\qquad$ ▷ See (4.8)

3: $\quad \mathbf{A}_{k+1} \leftarrow \arg\min_{\mathbf{A}} \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}_{k+1}\mathbf{A}^\top - \mathbf{S}_k\|_{\text{F}}^2 \quad \text{subject to} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I}$ $\quad$ ▷ See (4.9)

4: $\quad \mathbf{S}_{k+1} \leftarrow \arg\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}_{k+1}\mathbf{A}_{k+1}^\top - \mathbf{S}\|_{\text{F}}^2 + \kappa\|\mathbf{S}\|_1$ $\qquad$ ▷ See (4.8)

5: $\quad k \leftarrow k + 1$

6: **end while**

Output: $\mathbf{A}_k, \mathbf{B}_k, \mathbf{S}_k$

---

5. Spatiotemporal SPCA. Sparse decompositions are becoming increasingly relevant for data-driven spatiotemporal analysis of physical systems. The recent proliferation of machine learning and manifold learning methods seek interpretable models using physically meaningful constraints [42, 47, 33, 53, 35]. However, standard orthogonal decompositions such as SVD or proper orthogonal decomposition (POD) may suffer from overfitting and the resulting spatial modes are spatially dense. By promoting sparsity in the modes, SPCA is able to yield modes that may be more interpretable.

The goal of spatiotemporal modal analysis is a system decomposition that is separable in space and time,

$$(5.1) \qquad\qquad \mathbf{x}(t) = \sum_{j=1}^{r} a_j(t)\boldsymbol{\phi}_j,$$

where $\boldsymbol{\phi}_j$ is a mode evaluated at a grid of spatial locations. Classical data-driven analysis seeks a low-rank approximation given a data matrix of snapshots in time

$$(5.2) \qquad\qquad \mathbf{X} = [\mathbf{x}(t_1) \ \mathbf{x}(t_2) \ \ldots \ \mathbf{x}(t_p)].$$

The proper orthogonal decomposition is a canonical data-driven decomposition in the analysis of high-dimensional flows, which seeks the optimal rank-r orthogonal projection of the data that approximates the covariance of $\mathbf{X}$. The optimal low-rank projection is given by the dominant k scaled principal components $\mathbf{Z} = \mathbf{X}\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma}$, obtained from the singular value decomposition $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. We define the modes to be $\boldsymbol{\Phi} = \mathbf{U}$, resulting in the separable decomposition

$$(5.3) \qquad\qquad \mathbf{X} = \boldsymbol{\Phi}\mathbf{C}^{\mathbf{T}},$$

where $\mathbf{C}^{\mathbf{T}} = \boldsymbol{\Sigma}\mathbf{V}^T$. While POD modes numerically approximate the data, they may not be physically meaningful. POD modes do not generally correspond to coherent structures that persist in time. Sparse PCA, on the other hand, imposes sparsity in the spatial modes while maintaining time independence. In our framework, spatial modes are given by

$$(5.4) \qquad\qquad \boldsymbol{\Phi} = \mathbf{X}\mathbf{B} = \begin{bmatrix} | & | & & | \\ \mathbf{X}\mathbf{b}_1 & \mathbf{X}\mathbf{b}_2 & \ldots & \mathbf{X}\mathbf{b}_k \\ | & | & & | \end{bmatrix},$$

which represents a sparse linear combination of the snapshots and sparse modes $\boldsymbol{\Phi}$. Recall, the columns of $\mathbf{B}$ are the sparse weight vectors. As we shall demonstrate, SPCA modes display greater correspondence to coherent structures in various flows.

6. Results. We now apply our SPCA framework to a number of example systems of interest, ordered by increasing complexity. These examples capture many challenges that motivate the new algorithms. The first example is an artificial dataset with high-dimensional measurements and low-dimensional structures across multiple scales. In this example, there is a ground truth, providing a straightforward benchmark for SPCA and robust SPCA. The second example applies SPCA to a highly structured fluid flow, characterized by laminar vortex shedding behind a circular cylinder at low Reynolds number, which is a benchmark problem in fluid dynamics [41]. Fluid flows are ideal for developing interpretable models of multiscale physics and deploying sparse sensors for estimation and control. This is because they are high-dimensional systems that often exhibit low-dimensional coherent patterns that are spatially localized [9, 51]. The third example involves high-dimensional satellite data of the ocean surface temperature, a complex multiscale system that is intimately related to global circulation and climate. In all of these examples, the data is dynamic, high-dimensional, exhibits low-dimensional patterns at multiple scales, and has fewer snapshots in time than measurements in space. The proposed SPCA framework allows efficient computations on large systems, yields robust estimates from noisy data, and gives interpretable modes that can be used for the downstream tasks in dynamical systems modeling and control.

6.1. Multiscale Video Example. First, we consider a case where spatiotemporal dynamics are generated from three spatial modes oscillating at different frequencies in overlapping time intervals:

$$(6.1) \qquad \mathbf{x}(t) = \sum_{j=1}^{3} a_k(t)\boldsymbol{\phi}_j.$$

The multiscale time dynamics switch on and off irregularly, i.e., the modes effectively appear mixed in time as is common in other real-world phenomena such as weather, climate, etc. Consequently, within a single frame, the three modes occasionally mix, rendering the disambiguation task more challenging. However, we see that SPCA is able to recover the three modes in an unsupervised manner. Specifically, the data is generated on a 200×200 spatial grid for 150 seconds with timestep $\Delta t = .5$s. We flatten the spatial dimensions to obtain a data matrix with $p = 40,000$ measurements for each of the $n = 300$ snapshots (observations in time).

The results of PCA and SPCA on the raw frame data are compared and contrasted in Fig. 6. Here the spatial coherent structures extracted by SPCA recover the generating spatiotemporal modes, while PCA is unable to do so. By seeking a parsimonious representation, SPCA is able to accurately correlate spatial structures with their individual time histories. Because PCA has no such constraint, the different spatial structures remain mixed.

In many applications data exhibit grossly corrupted entries that typically arise from process or measurement noise. The least-squares loss function is sensitive to outliers. Thus, SPCA tends to be biased and the results can be misleading. To overcome this, our proposed robust SPCA algorithm can be used. The Huber loss function allows one to separate the input data into a low-rank component plus a sparse component. This is demonstrated in Fig. 7. The robust implementation clearly separates the polluted data into a low-rank component, while capturing the additive salt and pepper noise. However, the robust implementation is computationally more demanding than the standard SPCA algorithm.
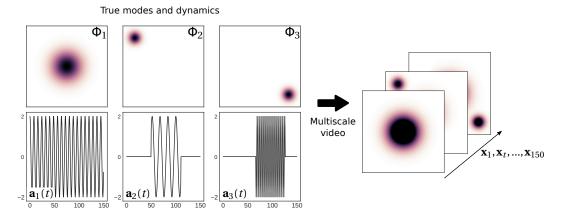
True modes and dynamics

Figure 5: Multiscale video model. Each frame of this multiscale video is high-dimensional with 200×200 pixels, however the system has only three degrees of freedom.

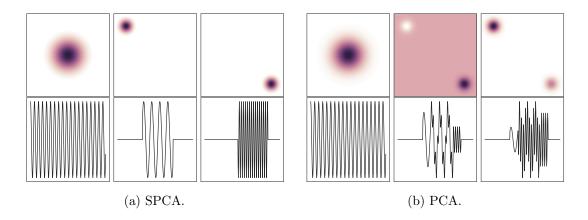(a) SPCA.   (b) PCA.

Figure 6: Multiscale video reconstruction. SPCA successfully decomposes the video into the true dynamics, while PCA fails to disambiguate modes 2 and 3.
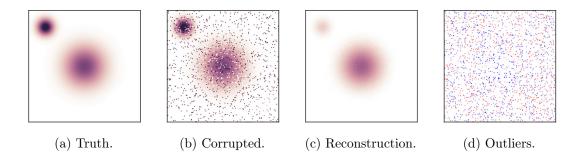

(a) Truth.   (b) Corrupted.   (c) Reconstruction.   (d) Outliers.

Figure 7: Approximation of a grossly corrupted multiscale video using robust SPCA. Here the low-rank approximation with robust PCA (c) successfully recovers the true frame and filters out added salt and pepper noise (d).

6.2. Fluid Flow. PCA has been extensively used in fluid dynamics for decades, where it is known as proper orthogonal decomposition, providing a data-driven generalization of the Fourier transform [6]. Here we apply SPCA to the flow behind a cylinder, a canonical example in fluid dynamics [41]. The data consists of a time series of the vorticity field behind a solid cylinder at Reynolds number 100, which induces laminar vortex shedding downstream. The flow is simulated using an immersed boundary projection method [52] on a 450×200 spatial grid for 3 dimensionless time units with timestep $\Delta t = .02$. Again, we flatten the spatial dimensions and yield a data matrix with $p = 90,000$ measurements for each of the $n = 150$ snapshots (observations in time). The resulting principal components or spatial modes of the flow are widely used for reduced-order modeling, prediction, and control.

The SPCA and PCA eigenmodes are compared in Fig. 8. Both decompositions successfully identify the dominant mode pairs that occur at characteristic harmonic frequencies. However, the mode structures extracted by SPCA are well-bounded and more interpretable, resulting in

(a) PCA.

(b) SPCA with ℓ_1 regularization ($\alpha = 10^{-5}$).

(c) SPCA with ℓ_1 regularization ($\alpha = 10^{-4}$).

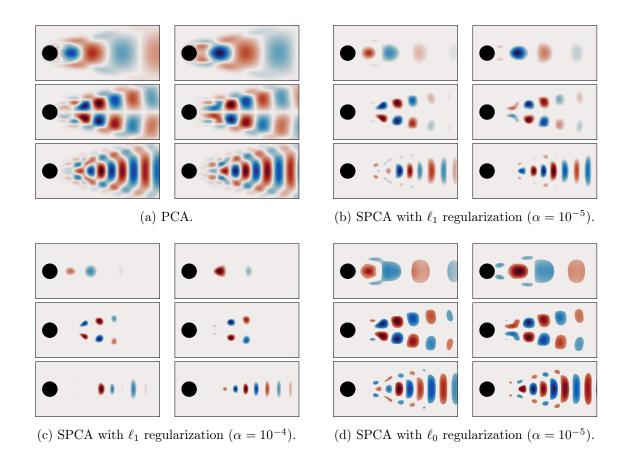(d) SPCA with ℓ_0 regularization ($\alpha = 10^{-5}$).

Figure 8: Sparse PCA demonstrates superior separation of the spatial eigenmodes responsible for vortex shedding. As a result, we can better differentiate their spatial influence on different regions of the flow downstream of the cylinder.

visible weakening downstream and stronger influence upstream. This is typical of the vortex shedding regime as vortices dissipate while advecting downstream and is not observed in the PCA modes.

Standard PCA has beneficial orthonormality properties that are crucial for projection based reduced-order modeling of high-dimensional systems. However, as experiments and models simulate increasingly complex flows, the field is rapidly moving towards more interpretable decompositions for learning and control. Recent directions in network analysis of turbulence and mixing require robust tracking of sparse spatial structures and vortices. The ability of SPCA to delineate boundaries of vortex dynamics are critical for the scalable decomposition of such high-resolution flow data. Furthermore, SPCA is purely data-driven and works equally well for modal decomposition of high-fidelity computational fluid dynamics (CFD) simulation, as well as robust denoising of experimental data generated by particle image velocimetry and other high-resolution imaging techniques.

6.3. NOAA Ocean Surface Temperature. We now apply SPCA to satellite ocean temperature data from 1990-2017 [45], and compare SPCA results from PCA.[1] The data consists of $n = 1,458$ temporal snapshots which measure the weekly temperature means at $360 \times 180 = 64,800$ spatial grid points. Since we omit data over continents and land, the ambient dimension reduces to $p = 44,219$ observations in our analysis. Our objective is the accurate identification of the intermittent El Niño and La Niña warming events, which are famously implicated in global weather patterns and climate change. The El Niño Southern Oscillation (ENSO) is defined as any sustained temperature anomaly above running mean temperature with a duration of 9 to 24 months. In climate sciences, principal components are also known as empirical orthogonal functions or EOFs; however, traditional PCA struggles to find a low-rank representation of this complex, high-dimensional system.

The canonical El Niño is associated with a narrow band of warm water off coastal Peru that is commonly referred as NIÑO 1+2, 3, 3.4, or 4 to differentiate the types of bands. Traditional PCA is unable to isolate this band, instead combining it with broader spatial signatures across the Pacific and Atlantic in mode 4 (Fig. 9a). Nevertheless, this mode is often used to compute the canonical Oceanic Niño Index (ONI). On the other hand, SPCA obtains a dramatic and clean separation of NIÑO 1-4 within the 4th mode (Fig. 9b). This is contextualized by the associated temporal mode, which yields sharper peaks during the 1997-1999 and 2014-2016 major El Niño events compared to PCA. The 12-month moving average of the temporal modes for both PCA and SPCA is shown in Fig. 10 and confirms that SPCA differentiates major and minor ENSO events with greater clarity than PCA.

Previous study of this dataset has required a multiresolution time-frequency separation of the data matrix in order to clearly identify the ENSO mode in an unsupervised manner [34].Without the sparsity constraint, SVD-based methods struggle to obtain a low-rank representation of these complex systems with nonlinear dynamics, coupled interactions and multiple timescales of motion. Based on our findings, SPCA has the potential to yield sparse modal representations of complex systems and coherent structures that may alter our understanding of oceanic and atmospheric phenomena.

---

[1]The data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, and are accessible via their Web site at https://www.esrl.noaa.gov/psd/.

(a) PCA.

(b) SPCA with ℓ_1 regularization ($\alpha = 10^{-4}$).

(c) SPCA with ℓ_0 regularization ($\alpha = 10^{-5}$).

(d) SPCA with ℓ_0 regularization ($\alpha = 10^{-4}$).
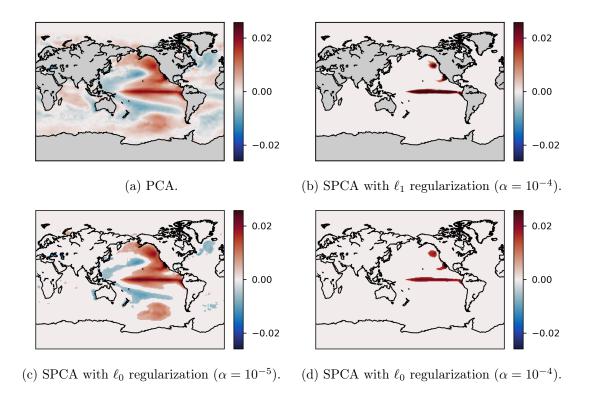
Figure 9: SPCA successfully identifies the band of warmer temperatures (4th mode) in the South Pacific traditionally associated with El Niño. By contrast, the corresponding PCA mode (4th mode) picks up spurious spatial correlations across the globe.
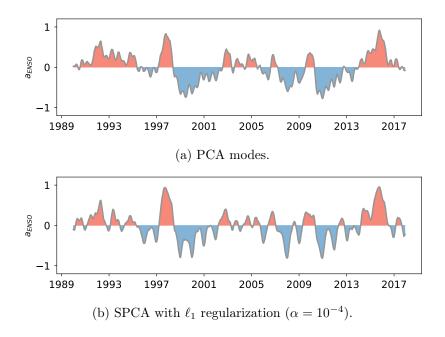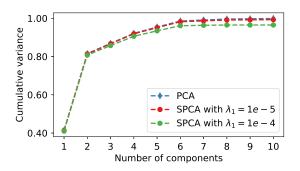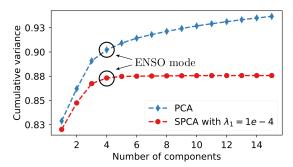

(a) PCA modes.

(b) SPCA with ℓ_1 regularization ($\alpha = 10^{-4}$).

Figure 10: Oceanic Niño Index (ONI), a 12-month moving average of the ENSO mode, reveals greater distinction between major (1997-1999,2014-2016) and minor events with SPCA modes.

6.4. Denoising with Sparse PCA. Cumulative variance plots reveal that sparse PCA behaves differently on the latter two examples. This can be attributed to the level of stochasticity in each system. The cylinder data has high temporal resolution and is therefore sufficiently well-resolved for sparse PCA to capture nearly all the variance within the low-rank component (Fig. 11a). In this case the decomposition is similar to PCA, although spatially more localized. On the other hand, the ocean data has coarse weekly temporal resolution. Therefore, faster dynamics which are not sufficiently resolved appear stochastic. Hence, slower timescales (annual, ENSO) are reflected in the low-rank component of SPCA as indicated by cumulative variance (Fig. 11b). PCA, however, overfits with 'noisy' components which are not physically meaningful.

6.5. Computational Performance. To demonstrate the computational performance of the proposed SPCA algorithms we compute the leading $k = 10$ components for two data matrices. First, we consider the cases of small p data. Figure 12 shows the time until the objective function converges within a tolerance level of 10^{-5}. For comparison we show the performance of the SPCA algorithm using least angle regression (LARS) and coordinate descent (CD) as proposed by [59]. Our proposed algorithm based on variable projection outperforms both the LARS and CD algorithm.

Further, the randomized accelerated SPCA algorithm outperforms the deterministic variable projection algorithms. The desired accuracy is achieved about 5 times faster compared to the deterministic algorithm. This is despite the fact that the randomized algorithms require more iterations than the deterministic algorithm to converge. The computational advantage is even greater for the high-dimensional data setting (i.e., big p) as shown in Figure 13 The computational advantage of the randomized algorithm becomes pronounced with increasing dimensions of the input matrix. Hence, the randomized algorithm allows exploring a large space of tuning parameters and is well suited for performing cross-validation.

(a) Although sparsity promotes spatially localized structure, SPCA retains nearly all the variance of the fluid flow system.

(b) SPCA clearly isolates the fourth ENSO mode as the last physically relevant component to the system.

Figure 11: Cumulative variance of each component. SPCA approximates PCA to varying degrees for the two fluid datasets. In contrast to PCA, SPCA separates the ENSO mode from noisy contributions even though ENSO captures only 1% of the total variance.

Implementations of our algorithms are provided in Python https://github.com/erichson/ristretto and in R https://CRAN.R-project.org/package=sparsepca.
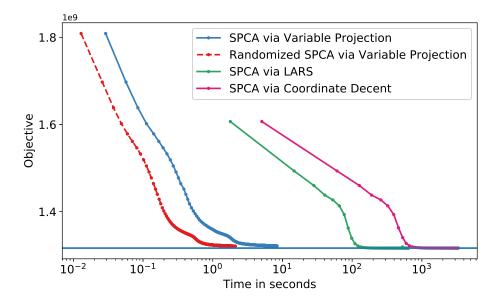

Figure 12: Computational performance of different SPCA algorithm. The dominant 10 sparse weight vectors are computed for a 2000×1344 data matrix.
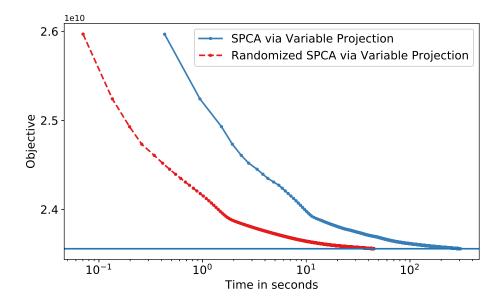

Figure 13: Computational performance of the randomized and deterministic SPCA algorithm using variable projection. The dominant 10 sparse weight vectors are computed for a 2000×16128 data matrix. The randomized algorithms is about 4 times faster.

7. Discussion. We have presented a robust and scalable architecture for computing sparse principal component analysis (SPCA). Specifically, we have modeled SPCA as a matrix factorization problem with orthogonality constraints, and developed specialized optimization algorithms that partially minimize a subset of the variables (variable projection). Our SPCA algorithm is scalable and robust, greatly improving computational efficiency over current state-of-the-art methods while retaining comparable performance. More precisely, we have demonstrated that: (i) The value function view approach provides an efficient and flexible framework for SPCA; (ii) Robust SPCA can be formulated using the Huber loss; (iii) A wide variety of sparsity-inducing regularizers can be incorporated into the framework; (iv) The proposed algorithms are computationally efficient for high-dimensional data, i.e, large p; (v) Randomized methods for linear algebra substantially eases the computational demands, while obtaining a near-optimal approximation for low-rank data.

SPCA is a useful diagnostic tool for data featuring rich dynamics that give rise to multiscale structures in both space and time. Given that such phenomena are ubiquitous in the physical, engineering, biological, and social sciences, this work provides a valuable tool for improved interpretability, especially in the diagnostics of localized structures and disambiguation of distinct time scale physical processes. The work also opens a number of avenues for future development:

Methodological Extensions. This scalable approach for identifying spatially localized spatial structures in high-dimensional and multiscale data may be directly applied to 1) tensor decompositions [14, 8, 21], which represent data in a multi-dimensional array structure, 2) parsimonious dynamical systems models [10], which identify the fewest nonlinear interactions required to capture the underlying physical mechanisms, and 3) *in situ* sensing and control, where sensors and actuators are generally required to be spatially localized [37].

Applications in the Engineering and Physical Sciences.. The methods developed here will be broadly applicable to dynamical systems that are high-dimensional, multiscale, and where there is a need for interpretable and parsimonious models for prediction, estimation, and control. Specific applications where SPCA has already been applied include genomics, and biological systems in general, and atmospheric chemistry. In addition, there is tremendous opportunity for advances in diverse fields, such as improving climate science, detecting and controlling structures in the brain, and closed loop control of turbulent fluid systems [9].

Appendix A. The Orthogonal Procrustes Problem.
We seek an orthonormal matrix $\mathbf{A}$ so that

$$(A.1) \qquad \mathbf{A} = \arg\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|_F^2 \quad \text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I}.$$

Indeed, a closed form solution is provided by the SVD. First,we expand the above objective

function as

$$\arg\min_{\mathbf{A}} \|\mathbf{X}\|_F^2 + \|\mathbf{X}\mathbf{B}\|_F^2 - 2 \cdot \text{trace}(\mathbf{X}^\top \mathbf{X}\mathbf{B}\mathbf{A}^\top).$$

This problem is equivalent to find a orthonormal matrix $\mathbf{A}$ which maximizes $\text{trace}(\mathbf{X}^\top \mathbf{X}\mathbf{B}\mathbf{A}^\top)$. We proceed by substituting the SVD of $\mathbf{X}^\top \mathbf{X}\mathbf{B}$ so that we yield

$$(A.2) \qquad \arg\max_{\mathbf{A}} \text{trace}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \mathbf{A}^\top) = \text{trace}(\boldsymbol{\Sigma}\mathbf{V}^\top \mathbf{A}^\top \mathbf{U}).$$

Note that $\boldsymbol{\Sigma}$ is a diagonal matrix with non-negative entries and $\mathbf{V}^\top \mathbf{A}^\top \mathbf{U}$ is an orthonormal matrix for any matrix $\mathbf{A}^\top$. Because of this, the trace norm in Eq. (A.2) is maximized by $\mathbf{A}^\top$ which turns $\mathbf{V}^\top \mathbf{A}^\top \mathbf{U}$ into an identity matrix $\mathbf{I}$, so that we yield $\text{trace}(\boldsymbol{\Sigma}\mathbf{I})$. Hence, an optimal solution is provided by $\mathbf{A} = \mathbf{U}\mathbf{V}^\top$, i.e., the left and right singular vectors of $\mathbf{X}^\top \mathbf{X}\mathbf{B}$.

Appendix B. Proof of Theorem.

B.1. Technical Preliminaries.
In the following we give a brief overview of notation and concepts used to develop and analyze the algorithms in this paper. Further, we review briefly the elements of variational analysis for the theoretical analysis of the algorithm [40, 46].

B.1.1. Matrix Spaces.
We consider the collection of all matrices with the same dimension $\mathbb{R}^d$ (where d could be shorthand for $p \times p$) as a Hilbert space equipped with the inner product. More concretely, the inner product is defined by the trace and the norm induced by this inner product is the Frobenius norm

$$\langle \mathbf{M}, \ \mathbf{M} \rangle := \text{trace}(\mathbf{M}^\top \mathbf{M}) = \|\mathbf{M}\|_F^2.$$

For any map $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^l$, we set,

$$\text{Lip}(\boldsymbol{\Phi}) := \sup_{\mathbf{M} \neq \mathbf{N}} \frac{\|\boldsymbol{\Phi}(\mathbf{M}) - \boldsymbol{\Phi}(\mathbf{N})\|_F}{\|\mathbf{M} - \mathbf{N}\|_F}$$

We say that $\boldsymbol{\Phi}$ is L-Lipschitz continuous, for some $L \geq 0$, if the inequality $\text{Lip}(\boldsymbol{\Phi}) \leq L$ holds.

B.1.2. Functions and Geometry.
Constraints, such as those in (3.1), can be represented using functions from a matrix space $\mathbb{R}^d$ to the extended real line defined by $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. The *domain* and the *epigraph* of any function $f : \mathbb{R}^d \to \mathbb{R}$ are the defined sets

$$\text{dom}\, f := \{\mathbf{M} \in \mathbb{R}^d : f(\mathbf{M}) < +\infty\},$$
$$\text{epi}\, f := \{(\mathbf{M}, r) \in \mathbb{R}^d \times \mathbb{R} : f(\mathbf{M}) \leq r\}.$$

For any set $\mathcal{F} \subset \mathbb{R}^d$, we define the *distance*, *projection* and *indicator function* for $\mathbf{M} \in \mathbb{R}^d$ by

$$\text{dist}(\mathbf{M}; \mathcal{F}) := \inf_{\mathbf{N} \in \mathcal{F}} \|\mathbf{N} - \mathbf{M}\|, \quad \text{proj}(\mathbf{M}; \mathcal{F}) := \arg\min_{\mathbf{N} \in \mathcal{F}} \|\mathbf{N} - \mathbf{M}\|,$$
$$\delta_{\mathcal{F}}(\mathbf{M}) := \begin{cases} 0, & \mathbf{M} \in \mathcal{F} \\ \infty, & \mathbf{M} \notin \mathcal{F} \end{cases}.$$

For $\mathbb{O} := \{\mathbf{A} \in \mathbb{R}^d : \mathbf{A}^T\mathbf{A} = \mathbf{I}\}$ in (3.1), and given $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, we have

$$\text{(B.1)} \qquad \text{dist}(\mathbf{M}; \mathbb{O}) := \|\mathbf{I} - \mathbf{\Sigma}\|^2, \quad \text{proj}(\mathbf{M}; \mathbb{O}) := \mathbf{U}\mathbf{V}^T, \quad \delta_{\mathbb{O}}(\mathbf{M}) := \begin{cases} 0, & \mathbf{\Sigma} = \mathbf{I} \\ \infty, & \mathbf{\Sigma} \neq \mathbf{I}. \end{cases}$$

B.1.3. Subgradients and Subdifferentials. Characterizing stationarity (a necessary condition for optimality) is a key step in analyzing the behavior of an algorithm and deriving practical termination criteria. Problem (3.1) is nonsmooth, so gradients do not exist. Instead, we can use more general concepts of *subgradients*, which exist for nonsmooth, nonconvex functions.

Consider an arbitrary function $f \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ and a point $\overline{\mathbf{M}}$ with $f(\overline{\mathbf{M}})$ finite. When f is convex, the subgradient of f at $\overline{\mathbf{M}}$ is defined as the collection of tangent affine minorants:

$$\text{(B.2)} \qquad \partial f(\overline{\mathbf{M}}) := \{\mathbf{V} : f(\mathbf{M}) \geq f(\overline{\mathbf{M}}) + \langle \mathbf{V}, \mathbf{M} - \overline{\mathbf{M}} \rangle\}.$$

If f is differentiable at $\overline{\mathbf{M}}$, then $\partial f(\overline{\mathbf{M}})$ contains only one element, and it is a gradient. When f is not differentiable, the subdifferential can contain multiple elements (see Figure 14). From (B.2), it is clear that $0 \in \partial f(\overline{\mathbf{M}})$ implies that $f(\mathbf{M}) \geq f(\overline{\mathbf{M}})$ for all $\mathbf{M}$, i.e. $\overline{\mathbf{M}}$ is a global minimum.

When f is nonconvex, (B.2) may not hold globally for any $\mathbf{V}$, and we need a localized definition. The *Fréchet subdifferential* of f at $\overline{\mathbf{M}}$, denoted $\hat{\partial}f(\overline{\mathbf{M}})$, is the set of all matrices $\mathbf{V}$ that satisfy

$$f(\mathbf{M}) \geq f(\overline{\mathbf{M}}) + \langle \mathbf{V}, \mathbf{M} - \overline{\mathbf{M}} \rangle + o(\|\mathbf{M} - \overline{\mathbf{M}}\|)$$

as $\mathbf{M} \to \overline{\mathbf{M}}$. The inclusion $\mathbf{V} \in \hat{\partial}f(\overline{\mathbf{M}})$ holds precisely when the affine function $\mathbf{M} \mapsto f(\overline{\mathbf{M}}) + \langle \mathbf{V}, \mathbf{M} - \overline{\mathbf{M}} \rangle$ underestimates f up to first-order near $\overline{\mathbf{M}}$. The limit of Fréchet subgradients $v_i \in \hat{\partial}f(\mathbf{M}_i)$ along a sequence $\mathbf{M}_i \to \overline{\mathbf{M}}$ may not be a Fréchet subgradient at the limiting point $\overline{\mathbf{M}}$. The *limiting subdifferential* $\partial f(\overline{\mathbf{M}})$ is the set of all matrices $\mathbf{V}$ for which there exist sequences $\mathbf{M}_i$ and $\mathbf{V}_i$ that satisfy $\mathbf{V}_i \in \partial f(\mathbf{M}_i)$ and $(\mathbf{M}_i, f(\mathbf{M}_i), \mathbf{V}_i) \to (\overline{\mathbf{M}}, f(\overline{\mathbf{M}}), \mathbf{V})$. In the nonconvex case, the stationarity condition $0 \in \partial f(\overline{\mathbf{M}})$ no longer implies global (or local) optimality. However, it is still a necessary condition, and one that can be checked. We characterize stationarity of (3.1) by the distance of 0 to the limiting subdifferential $\partial f(\overline{\mathbf{M}})$.

B.1.4. Moreau Envelope and Proximal Mapping. For any function f and real $\gamma > 0$, the *Moreau envelope* and the *proximal mapping* are defined by

$$\text{(B.3)} \qquad \begin{aligned} f_\gamma(\mathbf{M}) &:= \inf_{\mathbf{L}} \left\{ f(\mathbf{L}) + \frac{1}{2\gamma} \|\mathbf{L} - \mathbf{M}\|^2 \right\}, \\ \text{prox}_{\gamma f}(\mathbf{M}) &:= \arg\min_{\mathbf{L}} \left\{ f(\mathbf{L}) + \frac{1}{2\gamma} \|\mathbf{L} - \mathbf{M}\|^2 \right\}. \end{aligned}$$

Theorem B.1 (Regularization properties of the envelope). *Let* $f \colon \mathbb{R}^d \to \mathbb{R}$ *be a proper closed convex function. Then* f_γ *is convex and* C^1*-smooth with*

$$\nabla f_\gamma(\mathbf{M}) = \tfrac{1}{\gamma}(\mathbf{M} - \text{prox}_{\gamma f}(\mathbf{M})) \quad \text{and} \quad \text{Lip}(\nabla f_\gamma) \leq \tfrac{1}{\gamma}.$$

Proof. See Theorem 2.26 of [46]. ∎

(a) $f(x) = x^2$

(b) $f(x) = |x|$

(c) $f(x) = -|x|$

(d) $\partial(\cdot)^2(0) = \nabla(\cdot)^2(0) = 0.$

(e) $\partial|\cdot|(0) = [-1, 1].$
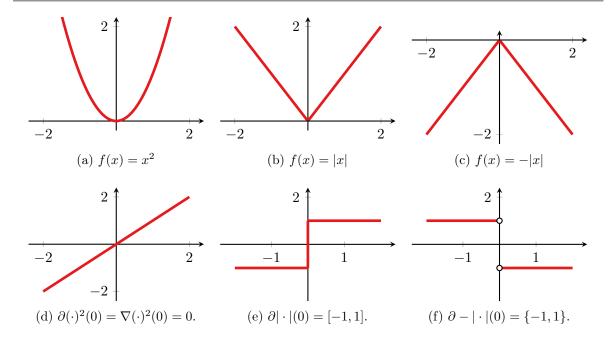
(f) $\partial - |\cdot|(0) = \{-1, 1\}.$

Figure 14: Subgradients are illustrated for the following three cases: (a) smooth function $f(x) = x^2$, (b) a nonsmooth function $f(x) = |x|$, (c) a nonsmooth and nonconvex function $f(x) = |x|$. Subplots (d) to (f) show the corresponding subgradients.

B.2. Optimality Condition. The objective function in Equation (3.3) is non-convex. Hence, we rely on iterative schemes that can find the stationary points of the objective.

Definition B.2 (Stationary Points). Assume that ρ is smooth, we call a pair $(\mathbf{A}, \mathbf{B})$ a stationary point when it satisfies

$$0 \in \nabla\rho(\mathbf{A}\mathbf{X}^\top\mathbf{B}^\top - \mathbf{X}^\top)\mathbf{B}\mathbf{X} + \partial\varphi(\mathbf{A}),$$
$$0 \in \mathbf{X}^\top\nabla\rho(\mathbf{X}\mathbf{B}\mathbf{A}^\top - \mathbf{X})\mathbf{A} + \partial\psi(\mathbf{B}),$$

where $\partial\psi$ is the limiting subdifferential defined in Section B.1. We also introduce a measure of non-stationarity T:

(B.4)
$$T(\mathbf{A}, \mathbf{B}) = \min\{\tfrac{1}{2}\|\mathbf{U}\|_{\mathrm{F}}^2 + \tfrac{1}{2}\|\mathbf{V}\|_{\mathrm{F}}^2 :$$
$$\mathbf{U} \in \nabla\rho(\mathbf{A}\mathbf{X}^\top\mathbf{B}^\top - \mathbf{X}^\top)\mathbf{B}\mathbf{X} + \partial\varphi(\mathbf{A}),$$
$$\mathbf{V} \in \mathbf{X}^\top\nabla\rho(\mathbf{X}\mathbf{B}\mathbf{A}^\top - \mathbf{X})\mathbf{A} + \partial\psi(\mathbf{B}).\}$$

In the main text, we show some examples of prox operator corresponding to $\partial\psi(\mathbf{B})$, but for $\partial\varphi(\mathbf{A})$ it might be hard to understand.

Here we give a simple instance of $\partial\varphi(\mathbf{A})$ when $\phi(\mathbf{A}) = \delta_0(\mathbf{A}|\mathbf{A}^\top\mathbf{A} = \mathbf{I}) = \delta_0(\mathbf{A}|\mathbb{O}_2)$, the space of orthonormal matrices. When consider orthogonal matrices in two dimension, we could

characterize them by a single angle variable θ,

$$\mathbf{A} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \text{ if } \det(\mathbf{A}) = 1, \quad \mathbf{A} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}, \text{ if } \det(\mathbf{A}) = -1.$$

For every $\mathbf{A}$ described as above, we define the tangent direction in $\mathbb{O}_2$:

$$\mathbf{T_A} = \begin{bmatrix} -\sin(\theta) & \cos(\theta) \\ -\cos(\theta) & -\sin(\theta) \end{bmatrix}, \text{ if } \det(\mathbf{A}) = 1, \quad \mathbf{T_A} = \begin{bmatrix} -\sin(\theta) & \cos(\theta) \\ \cos(\theta) & \sin(\theta) \end{bmatrix}, \text{ if } \det(\mathbf{A}) = -1.$$

We now have

$$\partial\varphi(\mathbf{A}) = \{\mathbf{G} : \langle \mathbf{G}, \ \mathbf{T_A} \rangle = 0\}.$$

In particular, for $\det(\mathbf{A}) = 1$ every element in $\partial\varphi(\mathbf{A})$ is a linear combination of the matrices

$$\begin{bmatrix} -\cos(\theta) & 0 \\ \sin(\theta) & 0 \end{bmatrix}, \quad \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}.$$

B.3. Proof for Theorem 4.1.

Proof. By definition, the iterates of Algorithm 4.1 satisfy

$$\frac{1}{\gamma}(\mathbf{B}_k - \mathbf{B}_{k+1}) + \partial\psi(\mathbf{B}_k) - \partial\psi(\mathbf{B}_{k+1}) \in \mathbf{X}^\top(\mathbf{X}\mathbf{B}_k - \mathbf{X}\mathbf{A}_k) + \partial\psi(\mathbf{B}_k)$$

$$\mathbf{0} \in (\mathbf{A}_{k+1}\mathbf{X}^\top\mathbf{B}_{k+1}^\top - \mathbf{X}^\top)\mathbf{B}_{k+1}\mathbf{X} + \partial\varphi(\mathbf{A}_{k+1}).$$

From the definition of the objective, we have,

$$\begin{aligned} f(\mathbf{A}_{k+1}, \mathbf{B}_{k+1}) &= \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}_{k+1}\mathbf{A}_{k+1}^\top\|_F^2 + \psi(\mathbf{B}_{k+1}) + \varphi(\mathbf{A}_{k+1}) \\ &\leq \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}_{k+1}\mathbf{A}_k^\top\|_F^2 + \varphi(\mathbf{A}_k) + \psi(\mathbf{B}_{k+1}) \\ &= \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}_k\mathbf{A}_k^\top + \mathbf{X}\mathbf{B}_k\mathbf{A}_k^\top - \mathbf{X}\mathbf{B}_{k+1}\mathbf{A}_k^\top\|_F^2 + \varphi(\mathbf{A}_k) + \psi(\mathbf{B}_{k+1}) \\ &= \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{B}_k\mathbf{A}_k^\top\|_F^2 + \langle \mathbf{X}(\mathbf{A}_k - \mathbf{B}_k), \ \mathbf{X}(\mathbf{B}_k - \mathbf{B}_{k+1}) \rangle \\ &\quad + \frac{1}{2}\|\mathbf{X}(\mathbf{B}_k - \mathbf{B}_{k+1})\|_F^2 + \psi(\mathbf{B}_{k+1}) + \varphi(\mathbf{A}_k) \\ &= f(\mathbf{A}_k, \mathbf{B}_k) + \langle \mathbf{X}(\mathbf{A}_k - \mathbf{B}_k), \ \mathbf{X}(\mathbf{B}_k - \mathbf{B}_{k+1}) \rangle \\ &\quad + \frac{1}{2}\|\mathbf{X}(\mathbf{B}_k - \mathbf{B}_{k+1})\|_F^2 + \psi(\mathbf{B}_{k+1}) - \psi(\mathbf{B}_k). \end{aligned}$$

Since ψ is a convex function, we have,

$$\psi(\mathbf{B}_{k+1}) - \psi(\mathbf{B}_k) \leq \langle \partial\psi(\mathbf{B}_{k+1}), \ \mathbf{B}_{k+1} - \mathbf{B}_k \rangle.$$

Therefore,

$$
\begin{aligned}
f(\mathbf{A}_{k+1}, \mathbf{B}_{k+1}) - f(\mathbf{A}_k, \mathbf{B}_k) \leq{}& \left\langle \mathbf{X}^\top \mathbf{X}(\mathbf{A}_k - \mathbf{B}_k),\ \mathbf{B}_k - \mathbf{B}_{k+1} \right\rangle + \frac{1}{2}\|\mathbf{X}(\mathbf{B}_k - \mathbf{B}_{k+1})\|_F^2 \\
&+ \left\langle \frac{1}{\gamma}(\mathbf{B}_k - \mathbf{B}_{k+1}) + \mathbf{X}^\top \mathbf{X}(\mathbf{A}_k - \mathbf{B}_k),\ \mathbf{B}_{k+1} - \mathbf{B}_k \right\rangle \\
={}& -\frac{1}{\gamma}\|\mathbf{B}_k - \mathbf{B}_{k+1}\|_F^2 + \frac{1}{2}\|\mathbf{X}(\mathbf{B}_k - \mathbf{B}_{k+1})\|_F^2 \\
\leq{}& -\frac{1}{2}\|\mathbf{X}\|_2^2 \|\mathbf{B}_k - \mathbf{B}_{k+1}\|_F^2.
\end{aligned}
$$

Using the definition of optimality condition T, we have

$$
\begin{aligned}
T(\mathbf{A}_k, \mathbf{B}_k) &\leq \left(\frac{1}{\|\mathbf{X}\|_2^2} + L \right)^2 \|\mathbf{B}_k - \mathbf{B}_{k+1}\|_F^2 \\
&\leq \frac{2(\|\mathbf{X}\|_2^2 + L)^2}{\|\mathbf{X}\|_2^2}(f(\mathbf{A}_k, \mathbf{B}_k) - f(\mathbf{A}_{k+1}, \mathbf{B}_{k+1}))
\end{aligned}
$$

Adding up the terms across k, we have a telescoping series on the right hand side, and immediately obtain the result. ∎

REFERENCES

[1] A. Aravkin and S. Becker, *Dual smoothing and value function techniques for variational matrix decomposition*, Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing, (2016), p. 2.

[2] A. Y. Aravkin, D. Drusvyatskiy, and T. van Leeuwen, *Efficient quadratic penalization through the partial minimization technique*, IEEE Transactions on Automatic Control, (2017).

[3] A. Y. Aravkin and T. van Leeuwen, *Estimating nuisance parameters in inverse problems*, Inverse Problems, 28 (2012), p. 115016.

[4] T. Askham and J. N. Kutz, *Variable projection methods for an optimized dynamic mode decomposition*, SIAM Journal on Applied Dynamical Systems, 17 (2018), pp. 380–416.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM journal on imaging sciences, 2 (2009), pp. 183–202.

[6] G. Berkooz, P. Holmes, and J. L. Lumley, *The proper orthogonal decomposition in the analysis of turbulent flows*, Annual review of fluid mechanics, 25 (1993), pp. 539–575.

[7] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, *Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset*, Computer Science Review, 23 (2017), pp. 1–71.

[8] R. Bro, *PARAFAC. Tutorial and applications*, Chemometrics and intelligent laboratory systems, 38 (1997), pp. 149–171.

[9] S. L. Brunton and B. R. Noack, *Closed-loop turbulence control: Progress and challenges*, Applied Mechanics Reviews, 67 (2015), pp. 050801–1–050801–48.

[10] S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proceedings of the National Academy of Sciences, 113 (2016), pp. 3932–3937.

[11] T. T. Cai, Z. Ma, and Y. Wu, *Sparse PCA: Optimal rates and adaptive estimation*, The Annals of Statistics, 41 (2013), pp. 3074–3110.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, Journal of the ACM (JACM), 58 (2011), p. 11.

[13] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-point algorithms for inverse problems in science and engineering, Springer, 2011, pp. 185–212.

[14] P. Comon, *Tensors: A brief introduction*, IEEE Signal Processing Magazine, 31 (2014), pp. 44–53.

[15] C. Croux, P. Filzmoser, and H. Fritz, *Robust sparse principal component analysis*, Technometrics, 55 (2013), pp. 202–214.

[16] J. P. Cunningham and Z. Ghahramani, *Linear dimensionality reduction: Survey, insights, and generalizations*, Journal of Machine Learning Research, 16 (2015), pp. 2859–2900.

[17] A. d'Aspremont, L. E. Ghaoui, M. I. Jordan, and G. R. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, in Advances in neural information processing systems, 2005, pp. 41–48.

[18] P. Drineas and M. W. Mahoney, *RandNLA: Randomized numerical linear algebra*, Communications of the ACM, 59 (2016), pp. 80–90.

[19] A. d'Aspremont, F. Bach, and L. E. Ghaoui, *Optimal solutions for sparse principal component analysis*, Journal of Machine Learning Research, 9 (2008), pp. 1269–1294.

[20] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al., *Least angle regression*, The Annals of statistics, 32 (2004), pp. 407–499.

[21] N. B. Erichson, K. Manohar, S. L. Brunton, and J. N. Kutz, *Randomized CP tensor decomposition*, arXiv preprint arXiv:1703.09074, (2017).

[22] N. B. Erichson, S. Voronin, S. L. Brunton, and J. N. Kutz, *Randomized matrix decompositions using R*, arXiv preprint arXiv:1608.02148, (2016).

[23] G. Golub and V. Pereyra, *Separable nonlinear least squares: The variable projection method and its applications*, Inverse Problems, 19 (2003), pp. R1–R26.

[24] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*, vol. 30, Oxford University Press, 2004.

[25] N. Halko, P.-G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM review, 53 (2011), pp. 217–288.

[26] J. H. Hong, C. Zach, and A. Fitzgibbon, *Revisiting the variable projection method for separable nonlinear least squares problems*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5939–5947.

[27] H. Hotelling, *Analysis of a complex of statistical variables into principal components.*, Journal of Educational Psychology, 24 (1933), p. 417.

[28] P. J. Huber, *Robust statistics*, in International Encyclopedia of Statistical Science, Springer, 2011, pp. 1248–1251.

[29] M. Hubert, T. Reynkens, E. Schmitt, and T. Verdonck, *Sparse PCA for high-dimensional data with outliers*, Technometrics, 58 (2016), pp. 424–434.

[30] I. T. Jolliffe, *Principal component analysis and factor analysis*, in Principal component analysis, Springer, 1986, pp. 115–128.

[31] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, *A modified principal component technique based on the lasso*, Journal of computational and Graphical Statistics, 12 (2003), pp. 531–547.

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

[33] E. Kaiser, B. R. Noack, L. Cordier, A. Spohn, M. Segond, M. Abel, G. Daviller, J. Östh, S. Krajnović, and R. K. Niven, *Cluster-based reduced-order modelling of a mixing layer*, Journal of Fluid Mechanics, 754 (2014), pp. 365–414.

[34] J. N. Kutz, X. Fu, and S. L. Brunton, *Multiresolution dynamic mode decomposition*, SIAM Journal on Applied Dynamical Systems, 15 (2016), pp. 713–735.

[35] J.-C. Loiseau and S. L. Brunton, *Constrained sparse galerkin regression*, Journal of Fluid Mechanics, 838 (2018), pp. 42–67.

[36] M. W. Mahoney, *Randomized algorithms for matrices and data*, Foundations and Trends in Machine Learning, 3 (2011), pp. 123–224.

[37] K. Manohar, B. W. Brunton, J. N. Kutz, and S. L. Brunton, *Data-driven sparse sensor placement for reconstruction*, IEEE Control Systems Magazine, 38 (2018), pp. 63–86.

[38] R. Maronna, R. D. Martin, and V. Yohai, *Robust statistics*, vol. 1, John Wiley & Sons, Chichester. ISBN, 2006.

[39] D. Meng, Q. Zhao, and Z. Xu, *Improve robustness of sparse PCA by l1-norm maximization*, Pattern

Recognition, 45 (2012), pp. 487 – 497.

[40] B. Mordukhovich, *Variational analysis and generalized differentiation. I*, vol. 330 of Grundlehren der Mathematischen Wissenschaften, Springer, 2006.

[41] B. R. Noack, K. Afanasiev, M. Morzynski, G. Tadmor, and F. Thiele, *A hierarchy of low-dimensional models for the transient and post-transient cylinder wake*, Journal of Fluid Mechanics, 497 (2003), pp. 335–363.

[42] V. Ozoliņš, R. Lai, R. Caflisch, and S. Osher, *Compressed modes for variational problems in mathematics and physics*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 18368–18373.

[43] N. Parikh, S. Boyd, et al., *Proximal algorithms*, Foundations and Trends® in Optimization, 1 (2014), pp. 127–239.

[44] K. Pearson, *On lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2 (1901), pp. 559–572.

[45] R. W. Reynolds, N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, *An improved in situ and satellite sst analysis for climate*, Journal of climate, 15 (2002), pp. 1609–1625.

[46] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317, Springer, 2009.

[47] H. Schaeffer, R. Caflisch, C. D. Hauck, and S. Osher, *Sparse dynamics for partial differential equations*, Proceedings of the National Academy of Sciences, 110 (2013), pp. 6634–6639.

[48] H. Shen and J. Z. Huang, *Sparse principal component analysis via regularized low rank matrix approximation*, Journal of multivariate analysis, 99 (2008), pp. 1015–1034.

[49] C. D. Sigg and J. M. Buhmann, *Expectation-maximization for sparse and non-negative PCA*, in Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 960–967.

[50] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, *A sparse-group lasso*, Journal of Computational and Graphical Statistics, 22 (2013), pp. 231–245.

[51] K. Taira, S. L. Brunton, S. Dawson, C. W. Rowley, T. Colonius, B. J. McKeon, O. T. Schmidt, S. Gordeyev, V. Theofilis, and L. S. Ukeiley, *Modal analysis of fluid flows: An overview*, AIAA Journal, 55 (2017), pp. 4013–4041.

[52] K. Taira and T. Colonius, *The immersed boundary method: A projection approach*, Journal of Computational Physics, 225 (2007), pp. 2118–2137.

[53] K. Taira, A. G. Nair, and S. L. Brunton, *Network structure of two-dimensional decaying isotropic turbulence*, Journal of Fluid Mechanics, 795 (2016), p. R2.

[54] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 91–108.

[55] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of optimization theory and applications, 109 (2001), pp. 475–494.

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

[57] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (2006), pp. 49–67.

[58] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 301–320.

[59] H. Zou, T. Hastie, and R. Tibshirani, *Sparse principal component analysis*, Journal of Computational and Graphical Statistics, 15 (2006), pp. 265–286.