SPARSE AND FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS

Genevera I. Allen

Departments of Statistics, CS, and ECE
Rice University
Houston, TX 77005
gallen@rice.edu

Michael Weylandt

Department of Statistics
Rice University
Houston, TX 77005
michael.weylandt@rice.edu

ABSTRACT

Regularized variants of Principal Components Analysis, especially Sparse PCA and Functional PCA, are among the most useful tools for the analysis of complex high-dimensional data. Many examples of massive data have both sparse and functional (smooth) aspects and may benefit from a regularization scheme that can capture both forms of structure. For example, in neuro-imaging data, the brain's response to a stimulus may be restricted to a discrete region of activation (spatial sparsity), while exhibiting a smooth response within that region. We propose a unified approach to regularized PCA which can induce both sparsity and smoothness in both the row and column principal components. Our framework generalizes much of the previous literature, with sparse, functional, two-way sparse, and two-way functional PCA all being special cases of our approach. Our method permits flexible combinations of sparsity and smoothness that lead to improvements in feature selection and signal recovery, as well as more interpretable PCA factors. We demonstrate the efficacy of our method on simulated data and a neuroimaging example on EEG data.

Index Terms— regularized PCA, multivariate analysis

1. INTRODUCTION

Principal Component Analysis (PCA) is a fundamental technique for dimension reduction, pattern recognition, and visualization of multivariate data. In the early 2000s, researchers noted that naive extensions of PCA to the high-dimensional setting produced unsatisfactory results, a finding later confirmed by advances in random matrix theory [1]. To address this limitation, many regularized variants of PCA were proposed, wherein the principal components were estimated under smoothness or sparsity assumptions [2]–[7]. Rather than reviewing this large literature, we instead refer the reader to the recent reviews of Hall [8], focusing on functional (smooth) PCA (FPCA) and of Zou and Xue [9], focusing on sparse PCA (SPCA).

Given the importance of both FPCA and SPCA, it is natural to ask whether it is possible to combine these approaches, yielding a unified approach to *sparse and functional PCA* (SFPCA). We show that this is indeed possible and present a unified optimization framework for doing so. Our proposed approach unifies much of the existing literature on regularized PCA; standard PCA, SPCA, FPCA, two-way SPCA, and two-way FPCA are all special cases of our approach, suggesting that it is, in some sense, the "correct" generalization.

Our unified SFPCA method enjoys many advantages over existing approaches to regularized PCA: i) because it allows for arbitrary degrees and forms of regularization, it is conducive to data-driven determination of the appropriate types and amount of regularization for a given problem; ii) because it unifies many existing methods, it inherits the desirable properties of both SPCA and FPCA, including superior signal recovery, automatic feature selection, and improved interpretability; and iii) it admits a tractable, efficient, and theoretically well-grounded algorithm.

Throughout this paper, we adopt the low-rank perspective on PCA and assume that our observed data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ arises from a low-rank structure $\boldsymbol{X} = \sum_{k=1}^{K} d_k \boldsymbol{u}_k \boldsymbol{v}_k^T + \boldsymbol{E}$, where the elements of $\boldsymbol{E}$ are independently and identically distributed with mean 0. We refer to the vectors $\{\boldsymbol{u}_k\}_{k=1}^K \in \mathbb{R}^n$ and $\{\boldsymbol{v}_k\}_{k=1}^K \in \mathbb{R}^p$ as the left and right singular vectors respectively. Given $\boldsymbol{X}$, its leading singular vectors can be estimated by solving the singular value problem:

$$\underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n, \boldsymbol{v} \in \overline{\mathbb{B}}^p}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} \tag{1}$$

where $\overline{\mathbb{B}}^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leq 1\}$ is the unit ball in $\mathbb{R}^n$. (Some authors require $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$, but, because the objective is linear in both $\boldsymbol{u}$ and $\boldsymbol{v}$, solutions to (1) lie on the boundary and this does not fundamentally change the problem.) Since the following singular vectors can be recovered by solving Problem (1) on a "deflated" $\boldsymbol{X}$, throughout this paper we principally focus on the leading singular vectors. Assuming that $\boldsymbol{X}$ has previously been centered, this approach is known to be equivalent to applying the eigenproblem formulation of PCA to both $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$.

2. A SPARSE AND FUNCTIONAL SINGULAR VALUE FORMULATION OF PCA

Taking the singular value problem (1) as a starting point, Huang *et al.* [4] proposed two-way FPCA by adding a product smoothness penalty

$$\underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n, \boldsymbol{v} \in \overline{\mathbb{B}}^p}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda \|\boldsymbol{u}\|_{\boldsymbol{S}_u}^2 \|\boldsymbol{v}\|_{\boldsymbol{S}_v}^2$$

where $\|\boldsymbol{u}\|_{\boldsymbol{S}_u}^2 = \boldsymbol{u}^T \boldsymbol{S}_u \boldsymbol{u}$ for some positive-definite $\boldsymbol{S}_u$ (similarly for $\boldsymbol{v}$). Typically, we take $\boldsymbol{S}_u = \boldsymbol{I} + \alpha_u \boldsymbol{\Omega}_u$ where $\boldsymbol{\Omega}_u$ is the second- or fourth-difference matrix, so that the $\|\boldsymbol{u}\|_{\boldsymbol{S}_u}^2$ penalty term encourages smoothness in the estimated singular vectors. Similarly, Allen *et al.* [7] proposed two-way SPCA by adding sparsity inducing penalties to the singular value problem (1):

$$\underset{\boldsymbol{u} \in \overline{\mathbb{B}}^n, \boldsymbol{v} \in \overline{\mathbb{B}}^p}{\arg\max} \; \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda_u P_u(\boldsymbol{u}) - \lambda_v P_v(\boldsymbol{v})$$

GA is also affiliated with the Jan and Dan Duncan Neurological Research Institute at the Baylor College of Medicine, Houston, TX 77030. GA acknowledges support from NSF DMS-1554821, NSF NeuroNex-1707400, and NSF DMS-126405. MW acknowledges support from the NSF Graduate Research Fellowship Program under grant number 1450681. The authors thank Dr. Yue Hu for assistance preparing the EEG Data used in Section 5 and Luofeng Liao for useful discussions.

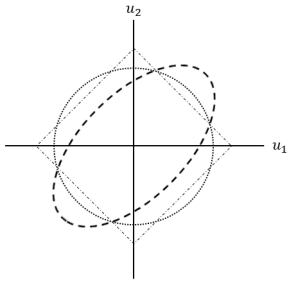

Fig. 1. Three constraints implicit in the ill-posed naive formulation of SFPCA: sparsity constraint (ℓ_1-ball), unit norm (ℓ_2-ball), and smoothness (elliptical region). In general, it is difficult for a point to lie on the boundary of all three regions simultaneously, leading to degenerate solutions to Problem (2).

where $P_{\boldsymbol{u}}$ and $P_{\boldsymbol{v}}$ are sparsity inducing penalties. (This is the Lagrangian form of the method of Witten *et al.* [5].) Given the success of these two methods, it is perhaps natural to perform SFPCA by adding both smoothness and sparsity penalties to Problem (1):

$$\underset{\boldsymbol{u}\in\overline{\mathbb{B}}^n,\boldsymbol{v}\in\overline{\mathbb{B}}^p}{\arg\max}\ \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v}-\lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u})-\lambda_{\boldsymbol{v}}P_{\boldsymbol{v}}(\boldsymbol{v})-\lambda\|\boldsymbol{u}\|^2_{\boldsymbol{S}_{\boldsymbol{u}}}\|\boldsymbol{v}\|^2_{\boldsymbol{S}_{\boldsymbol{v}}} \quad (2)$$

Surprisingly, this natural generalization fails, often spectacularly!

To see why this occurs, we note that Problem (2), with $\boldsymbol{v}$ held fixed, is actually attempting to satisfy three different constraints on $\boldsymbol{u}$ independently: a standard norm constraint, a smoothness constraint, and a sparsity constraint. As shown in Figure 1, unless all three regularization parameters ($\lambda, \alpha_{\boldsymbol{u}}, \lambda_{\boldsymbol{u}}$) are carefully chosen, this results in a form of "regularization masking," whereby it is impossible for the solution to Problem (2) to satisfy all constraints simultaneously. For the general case of two-way SFPCA, where we impose multiple constraints on both $\boldsymbol{u}$ and $\boldsymbol{v}$, this phenomenon is compounded.

To address the problem of regularization masking, we instead propose the following formulation of SFPCA:

$$\underset{\boldsymbol{u}\in\overline{\mathbb{B}}^n_{\boldsymbol{S}_{\boldsymbol{u}}},\boldsymbol{v}\in\overline{\mathbb{B}}^p_{\boldsymbol{S}_{\boldsymbol{v}}}}{\arg\max}\ \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v}-\lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u})-\lambda_{\boldsymbol{v}}P_{\boldsymbol{v}}(\boldsymbol{v}) \quad (3)$$

where $\overline{\mathbb{B}}^n_{\boldsymbol{S}_{\boldsymbol{u}}}$ is the unit ellipse of the $\boldsymbol{S}_{\boldsymbol{u}}$-norm, *i.e.*, $\overline{\mathbb{B}}^n_{\boldsymbol{S}_{\boldsymbol{u}}}=\{\boldsymbol{u}\in\mathbb{R}^n : \boldsymbol{u}^T\boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}\le 1\}$. As we will see below, this formulation is the "correct" generalization of many of the regularized PCA formulations previously proposed in the literature. Comparing our SFPCA formulation (3) with the naive formulation (2), we note two key differences: firstly, we only use a sparsity penalty in the objective function, moving the smoothness terms to the constraints to avoid regularization masking; secondly, we replace the unit ball constraint with a more general unit ellipse constraint. Since the unit ball constraint exists only to ensure identifiability of Problem (1), replacing it with a unit ellipse constraint simplifies the problem and ameliorates regularization masking. The benefits of this reformulation in eliminating regularization masking are formalized in Theorem 1 below.

Before proceeding, we make two regularity assumptions which we will use throughout our subsequent theoretical analysis:

Assumption 1. *In the SFPCA problem (3), with $\boldsymbol{S}_{\boldsymbol{u}}=\boldsymbol{I}+\alpha_{\boldsymbol{u}}\boldsymbol{\Omega}_{\boldsymbol{u}}$ and $\boldsymbol{S}_{\boldsymbol{v}}=\boldsymbol{I}+\alpha_{\boldsymbol{v}}\boldsymbol{\Omega}_{\boldsymbol{v}}$ for $\alpha_{\boldsymbol{u}},\alpha_{\boldsymbol{v}}\ge 0$, the following hold:*

(i) The smoothing matrices $\boldsymbol{\Omega}_{\boldsymbol{u}},\boldsymbol{\Omega}_{\boldsymbol{v}}$ are positive semi-definite.

(ii) The penalty terms $P_{\boldsymbol{u}},P_{\boldsymbol{v}}$ take values in $\mathbb{R}_{\ge 0}$ and are positive homogeneous of order one, i.e., *$P(c\boldsymbol{x})=cP(\boldsymbol{x})$ for all $c>0$ and all $\boldsymbol{x}$.*

Under these assumptions, our formulation of SFPCA (3) is well-posed and avoids many of the pathologies associated with other formulations:

Theorem 1. *Suppose Assumption 1 holds and let $(\boldsymbol{u}^*,\boldsymbol{v}^*)$ be the optimal points of the SFPCA problem (3). Then the following hold:*

(i) There exist values $\lambda^{\max}_{\boldsymbol{u}}$ and $\lambda^{\max}_{\boldsymbol{v}}$ such that, if $\lambda_{\boldsymbol{u}}\ge\lambda^{\max}_{\boldsymbol{u}}$ or if $\lambda_{\boldsymbol{v}}\ge\lambda^{\max}_{\boldsymbol{v}}$, then the solution to Problem (3) is trivial in the sense $(\boldsymbol{u}^*,\boldsymbol{v}^*)=(\boldsymbol{0},\boldsymbol{0})$.

(ii) If $\lambda_{\boldsymbol{u}}<\lambda^{\max}_{\boldsymbol{u}}$ and $\lambda_{\boldsymbol{v}}<\lambda^{\max}_{\boldsymbol{v}}$, the SFPCA solution $(\boldsymbol{u}^*,\boldsymbol{v}^*)$ depends on all (non-zero) regularization parameters.

(iii) $\|\boldsymbol{u}^*\|_{\boldsymbol{S}_{\boldsymbol{u}}}$ is equal to either 1 or 0, with the latter occurring only when $\lambda_{\boldsymbol{u}}\ge\lambda^{\max}_{\boldsymbol{u}}$ or $\lambda_{\boldsymbol{v}}\ge\lambda^{\max}_{\boldsymbol{v}}$. An analogous result holds for $\boldsymbol{v}^*$.

(iv) $(\boldsymbol{u}^*,\boldsymbol{v}^*)$ do not suffer from scale non-identifiability. (That is, $(c\boldsymbol{u}^*,c^{-1}\boldsymbol{v}^*)$ is not a solution for any $c\ge 0$ except $c=1$.)

The requirements of Assumption 1 are in fact quite weak and allow for nearly all the sparsity and smoothness structures previously proposed in the literature, including convex sparsity-inducing penalties (*e.g.*, the lasso [10]), structured-sparsity penalties such as the group or fused lasso [11], [12], and penalties based on the generalized lasso [13], as well as more exotic penalties such as the SLOPE penalty of Bogdan *et al.* [14]. As the following theorem shows, for various choices of the regularization parameters, SFPCA can yield the solution to standard PCA (SVD), SPCA, FPCA, two-way SPCA, and two-way FPCA:

Theorem 2. *Suppose Assumption 1 holds and let $(\boldsymbol{u}^*,\boldsymbol{v}^*)$ be the optimal points of the SFPCA problem (3). Then the following hold (up to a sign factor and unit scaling):*

(i) If $\lambda_{\boldsymbol{u}},\lambda_{\boldsymbol{v}},\alpha_{\boldsymbol{u}},\alpha_{\boldsymbol{v}}=0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are the first left and right singular vectors of $\boldsymbol{X}$.

(ii) If $\lambda_{\boldsymbol{u}},\alpha_{\boldsymbol{u}},\alpha_{\boldsymbol{v}}=0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the SPCA solution of Shen and Huang [15].

(iii) If $\alpha_{\boldsymbol{u}},\alpha_{\boldsymbol{v}}=0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the two-way SPCA solution in Allen et al. *[7], itself a special case of two-way sparse GPCA with the generalizing operators $\boldsymbol{Q},\boldsymbol{R}$ both identity matrices. (This is also the Lagrangian form of Witten* et al. *[5].)*

(iv) If $\lambda_{\boldsymbol{u}},\lambda_{\boldsymbol{v}},\alpha_{\boldsymbol{u}}=0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the FPCA solution of Silverman [2] and Huang et al. *[3].*

(v) If $\lambda_{\boldsymbol{u}},\lambda_{\boldsymbol{v}}=0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the two-way FPCA solution of Huang et al. *[4].*

For parts (ii) and (iii), equivalencies hold for the appropriate $P_{\boldsymbol{u}}(\cdot)$ and $P_{\boldsymbol{v}}(\cdot)$ employed in the referenced papers.

3. COMPUTATION OF SPARSE AND FUNCTIONAL PRINCIPAL COMPONENTS

We next present an efficient algorithm for computing sparse and functional components by solving Problem (3). The key to our algorithm is the observation that, if $P_{\boldsymbol{u}},P_{\boldsymbol{v}}$ are convex functions, then Problem (3) is a bi-concave problem in $\boldsymbol{u}$ and in $\boldsymbol{v}$, where each subproblem is equivalent to a penalized regression problem. This suggests an alternating proximal gradient ascent strategy, which yields the following rank-one SFPCA Algorithm, where $\lambda_{\max}(\boldsymbol{A})$ is the leading eigenvalue of $\boldsymbol{A}$ and $\mathsf{prox}_{f(\cdot)}(\boldsymbol{z})=\arg\min_{\boldsymbol{x}}\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2_2+f(\boldsymbol{x})$ is the proximal operator of f:

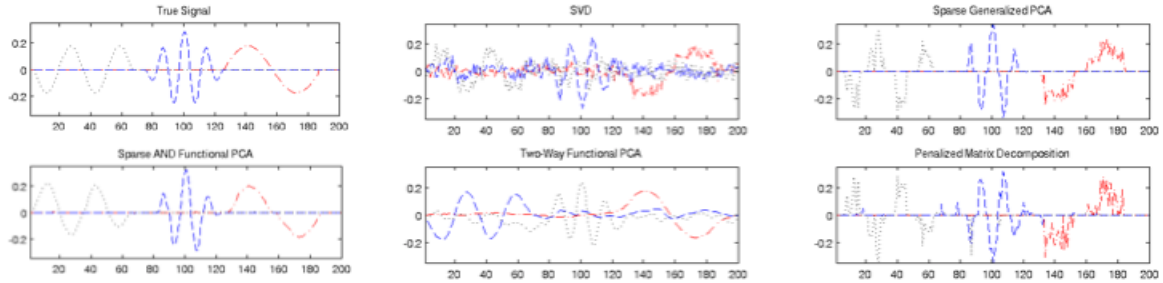

Fig. 2. Simulated factors used for the simulation study in Section 4 and estimates thereof: $\boldsymbol{v}_1$ (red, dotted-dashed); $\boldsymbol{v}_2$ (blue, dashed); and $\boldsymbol{v}_3$ (black, dotted). Only SFPCA is able to simultaneously identify the spatial sparsity and smooth structure of the sinusoidal pulses.

Algorithm 1 Rank-1 SFPCA Algorithm (Proximal Gradient Variant)

1. Initialize $\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}$ to the leading singular vectors of $\boldsymbol{X}$ and set $L_{\boldsymbol{u}} = \lambda_{\max}(\boldsymbol{S}_{\boldsymbol{u}})$ and $L_{\boldsymbol{v}} = \lambda_{\max}(\boldsymbol{S}_{\boldsymbol{v}})$
2. Repeat until convergence:
   (a) $\boldsymbol{u}$-subproblem: repeat until convergence:
   $$\boldsymbol{u} := \text{prox}_{\frac{\lambda_{\boldsymbol{u}}}{L_{\boldsymbol{u}}} P_{\boldsymbol{u}}(\cdot)} \left(\boldsymbol{u} + L_{\boldsymbol{u}}^{-1} \left(\boldsymbol{X}\hat{\boldsymbol{v}} - \boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}\right)\right)$$
   $$\hat{\boldsymbol{u}} := \begin{cases} \boldsymbol{u} & \|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} \leq 1 \\ \boldsymbol{u}/\|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} & \text{otherwise} \end{cases}$$
   (b) $\boldsymbol{v}$-subproblem: repeat until convergence:
   $$\boldsymbol{v} := \text{prox}_{\frac{\lambda_{\boldsymbol{v}}}{L_{\boldsymbol{v}}} P_{\boldsymbol{v}}(\cdot)} \left(\boldsymbol{v} + L_{\boldsymbol{v}}^{-1} \left(\boldsymbol{X}^T\hat{\boldsymbol{u}} - \boldsymbol{S}_{\boldsymbol{v}}\boldsymbol{v}\right)\right)$$
   $$\hat{\boldsymbol{v}} := \begin{cases} \boldsymbol{v} & \|\boldsymbol{v}\|_{\boldsymbol{S}_{\boldsymbol{v}}} \leq 1 \\ \boldsymbol{v}/\|\boldsymbol{v}\|_{\boldsymbol{S}_{\boldsymbol{v}}} & \text{otherwise} \end{cases}$$
3. Return $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$, optionally scaled to have (Euclidean) norm 1

In the final step, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ may be rescaled to have unit norm, as with standard PCA and other regularized variants, but if so, they may no longer be feasible for Problem (3). Despite the non-convexity of the SFPCA problem (3), Algorithm 1 comes with the following strong convergence guarantees:

Theorem 3. *Under Assumption 1, Algorithm 1 has the following properties:*

(i) Step 2(a) converges to a stationary point of

$$\underset{\boldsymbol{u} \in \mathbb{R}^n_{\boldsymbol{S}_{\boldsymbol{u}}}}{\arg\min} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{v} - \boldsymbol{u}\|_2^2 + \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) + \frac{\alpha_{\boldsymbol{u}}}{2}\boldsymbol{u}^T\boldsymbol{\Omega}_{\boldsymbol{u}}\boldsymbol{u}. \quad (4)$$

Furthermore, if $P_{\boldsymbol{u}}$ is convex, the convergence is monotone, at an $\mathcal{O}(1/K)$ rate, and to a global solution. Step 2(b) converges analogously for $\boldsymbol{v}$ and $P_{\boldsymbol{v}}$.

(ii) If $P_{\boldsymbol{u}}$ is convex, Step 2(a) yields a global solution to (3), considering $\hat{\boldsymbol{v}}$ fixed; if $P_{\boldsymbol{u}}$ is non-convex, Step 2(a) yields a stationary point for $P_{\boldsymbol{u}}$, considering $\hat{\boldsymbol{v}}$ fixed. An analogous result holds for $\hat{\boldsymbol{v}}$ returned by Step 2(b), with $\hat{\boldsymbol{u}}$ considered fixed.

(iii) If $P_{\boldsymbol{u}}, P_{\boldsymbol{v}}$ are both convex, then $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ returned by the SFPCA Algorithm (1) is both a coordinate-wise global maximum (Nash point) and a stationary point of Problem (3).

We note that the convergence rates associated with steps 2(a) and 2(b) can be further improved to $\mathcal{O}(1/K^2)$ if an accelerated proximal gradient scheme is instead used to solve the $\boldsymbol{u}$- or $\boldsymbol{v}$-subproblems [16], though monotonicity may be lost. Additionally, in the case where $\alpha_{\boldsymbol{u}} = 0$, then subproblem (4) is solved by normalizing $\text{prox}_{\lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\cdot)}(\boldsymbol{X}\boldsymbol{v})$ and hence converges in a single step.

Since the SFPCA problem (3) is non-convex, the estimates returned by Algorithm 1 depend on the initial values chosen for $\boldsymbol{u}$ and $\boldsymbol{v}$. In practice, we have found the unregularized singular vectors to provide a robust and easily computed initialization. More complex constraints can be added to SFPCA by incorporating them in the proximal operators applied in steps 2(a) and 2(b) of Algorithm 1. In particular, we can impose non-negativity constraints of the form considered by Allen and Maletić-Savatić [6] by incorporating the indicator function of the positive orthant into the penalty functions $P_{\boldsymbol{u}}, P_{\boldsymbol{v}}$; for many popular penalties, this yields a positive proximal operator with a closed form, *e.g.*, the positive-part operator when the underlying penalty is the lasso.

Algorithm 1 returns estimates of the leading left and right regularized singular vectors of $\boldsymbol{X}$ only. Additional regularized singular vectors can be obtained by iteratively applying Algorithm 1 to a suitably deflated data matrix. In our simulation and case studies in the next two sections, we use Hotelling's subtraction deflation ($\boldsymbol{X} := \boldsymbol{X} - d\hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^T$ where $d = \hat{\boldsymbol{u}}^T\boldsymbol{X}\hat{\boldsymbol{v}}$), though the alternative deflation schemes proposed by Mackey [17] could be also be used.

Because Algorithm 1 essentially only requires solving penalized regression problems, it avoids the expensive matrix inversion or eigendecomposition steps common to other regularized PCA variants. For problems with closed-form proximal operators that can be evaluated in linear time, the computational cost of Algorithm 1 is $\mathcal{O}(n^2 + p^2)$, dominated by the cost of multiplication by $\boldsymbol{S}_{\boldsymbol{u}}$ and $\boldsymbol{S}_{\boldsymbol{v}}$. As smoothing matrices typically have a banded structure, additional problem-specific improvements are often possible. We also note that randomized methods [18] can be used to efficiently obtain estimates of the leading singular vectors of $\boldsymbol{X}$ used to initialize $\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}$ in Algorithm 1, thereby avoiding an expensive computation in very large problems.

3.1. Selection of Regularization Parameters

While Algorithm 1 provides an efficient and scalable approach to fitting SFPCA on large data sets, we have not yet addressed the question of tuning various regularization parameters. The presence of four independently chosen tuning parameters – $\lambda_{\boldsymbol{u}}, \lambda_{\boldsymbol{v}}, \alpha_{\boldsymbol{u}}, \alpha_{\boldsymbol{v}}$ – would appear to be a major drawback of our formulation. Indeed, cross-validation over a four dimensional grid of regularization parameters would pose a significant computational burden. Instead we adapt the strategy of Huang *et al.* [4], who exploit the connection between two-way FPCA and penalized regression methods to develop an efficient tuning scheme.

In particular, we propose a greedy "coordinate-wise" Bayesian Information Criterion (BIC) optimization scheme. We begin by holding the tuning parameters associated with $\boldsymbol{v}$ fixed ($\alpha_{\boldsymbol{v}}, \lambda_{\boldsymbol{v}}$) and choosing $\alpha_{\boldsymbol{u}}$ and $\lambda_{\boldsymbol{u}}$ to optimize the BIC of the $\boldsymbol{u}$-subproblem (4). We then

			TWFPCA	SSVD	PMD	SGPCA ($\sigma = 1$)	SGPCA ($\sigma = 5$)	SFPCA
$n = 100$	$\boldsymbol{v}_1$	TP	-	0.897 (.004)	0.568 (.005)	0.768 (.008)	0.820 (.004)	**0.935** (.004)
		FP	-	0.323 (.080)	0.001 (.000)	**0.006** (.002)	0.012 (.002)	0.052 (.032)
		r$\angle$	**0.153** (.055)	0.625 (.112)	2.220 (.035)	0.726 (.024)	0.369 (.007)	0.189 (.062)
	$\boldsymbol{v}_2$	TP	-	**0.783** (.007)	0.657 (.006)	0.445 (.010)	0.005 (.002)	0.713 (.008)
		FP	-	0.320 (.080)	0.106 (.004)	**0.002** (.001)	0.257 (.003)	0.047 (.031)
		r$\angle$	5.980 (.346)	0.549 (.105)	0.597 (.012)	0.829 (.024)	6.150 (.104)	**0.438** (.094)
	$\boldsymbol{v}_3$	TP	-	0.771 (.007)	0.514 (.007)	0.499 (.015)	0.064 (.014)	**0.883** (.008)
		FP	-	0.316 (.079)	0.066 (.004)	**0.004** (.002)	0.128 (.014)	0.054 (.033)
		r$\angle$	3.660 (.270)	0.855 (.131)	1.270 (.023)	1.010 (.038)	4.000 (.093)	**0.468** (.097)
		rSE	0.668 (.003)	0.760 (.002)	1.000 (.008)	0.737 (.009)	0.936 (.017)	**0.450** (.003)
$n = 300$	$\boldsymbol{v}_1$	TP	-	0.973 (.002)	0.509 (.003)	0.921 (.003)	0.904 (.002)	**0.987** (.001)
		FP	-	0.322 (.080)	**0.000** (.000)	0.005 (.002)	0.015 (.002)	0.068 (.037)
		r$\angle$	0.768 (.124)	0.487 (.099)	15.700 (.292)	0.553 (.017)	0.443 (.011)	**0.152** (.055)
	$\boldsymbol{v}_2$	TP	-	0.919 (.004)	0.773 (.003)	0.839 (.004)	0.011 (.003)	**0.967** (.003)
		FP	-	0.319 (.080)	**0.000** (.000)	0.038 (.003)	0.323 (.002)	0.048 (.031)
		r$\angle$	52.300 (1.02)	0.428 (.093)	1.310 (.023)	0.488 (.024)	52.800 (.935)	**0.320** (.080)
	$\boldsymbol{v}_3$	TP	-	0.943 (.003)	0.530 (.004)	0.849 (.006)	0.005 (.002)	**0.972** (.002)
		FP	-	0.314 (.079)	**0.000** (.000)	0.015 (.003)	0.212 (.002)	0.060 (.035)
		r$\angle$	33.100 (.813)	0.545 (.104)	5.940 (.089)	0.631 (.026)	34.200 (.543)	**0.131** (.051)
		rSE	1.170 (.002)	0.790 (.001)	3.380 (.016)	0.809 (.005)	1.360 (.007)	**0.655** (.001)

Table 1. Performance of various regularized PCA methods for the simulation study described in Section 4. Results are averaged over 50 replicates, with standard errors given in parentheses. For each method, the true positive rate (TP), false positive rate (FP), relative angle compared to that of the SVD (r$\angle$), and relative squared error compared to that of the SVD (rSE) are reported. (TP and FP are not reported for the non-sparse TWFPCA.) The best performing method on each metric is bold-faced. SFPCA consistently outperforms other methods.

hold $\alpha_{\boldsymbol{u}}$ and $\lambda_{\boldsymbol{u}}$ and optimize the BIC of the $\boldsymbol{v}$-subproblem. If these searches are embedded within a warm-starting scheme for steps 2(a) and 2(b) of Algorithm 1, this can be achieved with minimal additional computational cost. The degrees of freedom and associated BIC of the $\boldsymbol{u}$- and $\boldsymbol{v}$-subproblems can be established using the techniques proposed by Kato [19] and Tibshirani and Taylor [20], though we provide an explicit expression for the common case of an ℓ_1 sparsity penalty:

Theorem 4. *Suppose $P_{\boldsymbol{u}}(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$. Then an unbiased estimate degrees of freedom of the $\boldsymbol{u}$-subproblem (4) is given by*

$$\widehat{\mathrm{df}}(\hat{\boldsymbol{u}}) = \mathrm{Tr}\left[\left(\boldsymbol{I}_{|\mathcal{A}|} + \alpha_{\boldsymbol{u}}\boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}}\right)^{-1}\right] \approx \mathrm{Tr}\left[\boldsymbol{I}_{|\mathcal{A}|} - \alpha_{\boldsymbol{u}}\boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}}\right]$$

where $\mathcal{A}$ denotes the indices of the estimated non-zero elements of $\hat{\boldsymbol{u}}$ and $\boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}}$ denotes the corresponding submatrix of $\boldsymbol{\Omega}_{\boldsymbol{u}}$. Hence, the approximate BIC to be optimized for subproblem (4) is given by

$$\mathrm{BIC}(\hat{\boldsymbol{u}}) = \log\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{v} - \hat{\boldsymbol{u}}\|_2^2\right] + \frac{1}{n}\log(n)\,\widehat{\mathrm{df}}(\hat{\boldsymbol{u}}).$$

One potential shortcoming of our proposed approach is that the greedy search is not guaranteed to converge and may enter an infinite loop as it attempts to optimize the regularization parameters. To address this, non-convergence guards (*e.g.*, a maximum number of steps) may be added, but in our experience, however, the greedy search tends to stabilize quickly and guards against non-convergence are not needed for most problems. As shown in the next two section, this scheme performs well in practice, selecting flexible combinations of sparsity and smoothness in a tractable data-driven manner.

4. SIMULATION STUDY

In this section, we compare the performance of our SFPCA method (3) with several competitors including the two-way FPCA (TWFPCA) method of Huang *et al.* [4], the sparse SVD method (SSVD) of Lee *et al.* [21], the penalized matrix decomposition (PMD) of Witten *et al.* [5], and the sparse generalized PCA (SGPCA) of Allen *et al.* [7]. We simulate data according to the low-rank model $\boldsymbol{X} = \sum_{k=1}^{K} d_k \boldsymbol{u}_k \boldsymbol{v}_k^T + \boldsymbol{E}$ where $E_{ij} \overset{\text{IID}}{\sim} \mathcal{N}(0,1)$. We fix $K = 3$ and $p = 200$ and sample the left singular vectors uniformly $\boldsymbol{v}$ from the space of orthogonal matrices. The signal in the right singular vectors $\boldsymbol{v}$, each of which have a combination of sparsity and smoothness, takes the form of a sinusoidal pulse. The scale-factors d_i, which control the signal-to-noise ratio, vary with the sample size as $d_1 = n/4, d_2 = n/5, d_3 = n/6$.

The SGPCA generalizing operators were constructed using the method suggested by Allen and Maletić-Savatić [6] with kernel $e^{-d_{ij}^2/\sigma}$ for Chebychev distances between time points i, j. The smoothing matrices $\Omega_{\boldsymbol{u}}, \Omega_{\boldsymbol{v}}$ were fixed as squared second difference matrices. The sparse methods were implemented using an unweighted ℓ_1-penalty. Tuning parameters for each method were selected using the authors' recommended approach. For SFPCA, the greedy BIC method described above was used.

Our qualitiative results are shown in Figure 2, where we see that SFPCA clearly outperforms the competing methods. The non-sparse standard SVD and TWFPCA are not able to successfully localize the sinusoidal pulses in time, while the non-smooth PMD and SGPCA are not able to recover the smooth sinusoidal structure.

Quantitative results are presented in Table 1, where we report the true positive rate (TP) and false positive rate (FP) for recovering the support of $\boldsymbol{v}$, as well as two measures of smoothness, the relative angle and the relative squared error. The relative angle is given by r$\angle = (1 - |\hat{\boldsymbol{v}}^T\boldsymbol{v}^*|)/(1 - |\hat{\boldsymbol{v}}_{\text{SVD}}^T\boldsymbol{v}^*|)$ where $\boldsymbol{v}^*$ is the true signal and $\hat{\boldsymbol{v}}_{\text{SVD}}$ is the SVD-estimated singular vector; smaller values of r$\angle$ indicate better performance, with values less than one signifying more accurate estimation than the standard SVD. The relative squared error measures the reconstruction accuracy and is given by $\text{rSE} = \|\boldsymbol{X}^* - \hat{\boldsymbol{X}}\|_F^2 / \|\boldsymbol{X}^* - \hat{\boldsymbol{X}}^{\text{3-SVD}}\|_F^2$; smaller values of rSE indicate better performance, with values less than one signifying more accurate estimation than the standard SVD. (Note that that for both measures, we consider reconstruction of the true mean matrix and true right singular vectors, else it would be impossible to outperform the SVD.) SFPCA consistently outperforms the other regularized

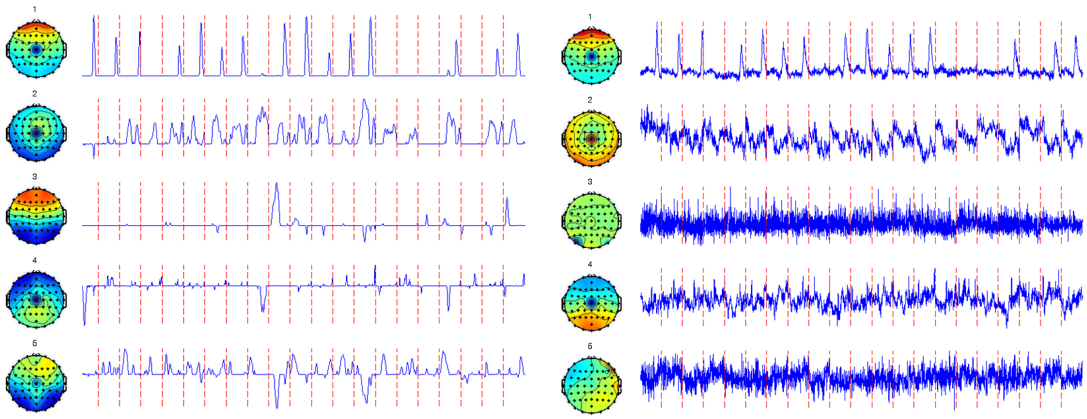

Fig. 3. EEG Case Study: first five spatial and temporal SFPCA components (left) and ICA components (right). While SFPCA and ICA identify similar structures in the first two components, the temporal sparsity of the SFPCA components makes them more readily interpretable. Additionally, the SFPCA finds structure in the subsequent components that ICA does not identify.

PCA methods and, as measured by r$\angle$ and rSE, the standard SVD. Clearly, SFPCA is able to accurately and adaptively recover principal components with complex structure, yielding improved statistical performance. As we will see in the next section, the structured principal components yielded by SFPCA are also more interpretable, making SFPCA a useful tool for exploratory data analysis and scientific model construction.

5. CASE STUDY: EEG DATA

We close with an application of SFPCA to a sample of electoencephalography (EEG) data taken from the UCI Machine Learning Repository [22].[1] These data consist of $n = 57$ EEG channels with corresponding scalp locations and $p = 5376$ time points, corresponding to 21 epochs of 256 time points each. Back-block pattern recognition techniques, especially independent components analysis (ICA), are commonly applied to EEG data to separate sources from the limited channel recordings, find major spatial patterns and corresponding temporal activity patterns, find artifacts in the data, and develop visualizations [23]. SFPCA was applied to the EEG recording from the first alcoholic subject over epochs relating to non-matching stimuli. The spatial smoothing matrix, $\mathbf{\Omega}_{\boldsymbol{u}}$, was specified as the weighted squared second differences matrix using spherical distances between the recording channel locations and the temporal smoothing matrix, $\mathbf{\Omega}_{\boldsymbol{v}}$, was taken as the matrix of squared second differences. Tuning parameters for SFPCA were selected using the greedy scheme described above.

In Figure 3, we compare the SFPCA results with those obtained from the FastICA method [24]. At a high level, the patterns identified by SFPCA and ICA are similar, identifying the same major temporal patterns and spatial source localization, but the SFPCA results are much more directly interpretable. The improvements afforded by SFPCA are clearly seen by comparing the first two components, where the spatial patterns are similar but SFPCA identifies a much more structured temporal pattern. Furthermore, SFPCA is able to identify more signals: the third SFPCA vectors identify a singular "pulse" which is spatially and temporally localized, while the third ICA component has no discernable structure.

[1] https://archive.ics.uci.edu/ml/datasets/eeg+database

Interestingly, the greedy BIC scheme consistently selects $\lambda_{\boldsymbol{u}} = 0$, suggesting that no sparsity in the EEG channels is needed. Conversely, the greedy scheme consistently selected non-zero smoothing and temporal sparsity parameters for each of the first five SFPCA components ($\alpha_{\boldsymbol{u}} \in [10, 12]$, $\alpha_{\boldsymbol{v}} \in [0.5, 10]$, $\lambda_{\boldsymbol{v}} \in [1, 2.5]$), indicating that our method is able to flexibly choose the optimal degree of smoothness and sparsity for recovering major patterns in the data.

6. DISCUSSION

We have proposed SFPCA, a flexible yet coherent approach to sparsity- and smoothness-regularized PCA. This flexibility gives SFPCA the ability to adapt to the types and amounts of regularization appropriate for a given problem in a data-driven manner. SFPCA unifies much of the existing literature on regularized PCA and allows for as-of-yet-unexplored generalizations by varying the penalty functions and smoothing matrices. In our simulation and case studies, SFPCA exhibits superior statistical performance and improved interpretability. As special cases of SFPCA have been shown to lead to consistent estimation of principal components, even in the high-dimensional context [2], [25], we conjecture that the general SFPCA framework also yields consistent estimates, an interesting topic for future research.

The advantages of SFPCA are not purely theoretical, however: Algorithm 1 provides a framework for solving the SFPCA Problem, which is fast and scalable for general problems, while also easily modified to take advantage of additional computational efficiencies afforded by specific problems. As shown in Theorem 3, Algorithm 1 enjoys attractive convergence properties despite its inherent non-convexity. Additionally, the greedy BIC scheme we have proposed allows for computationally efficient determination of regularization parameters. MATLAB scripts implementing SFPCA are available from the first author's website. Supplemental materials for this paper including proofs and additional experiments are available at https://arxiv.org/abs/1309.2895.

The advantages of SFPCA demonstrated here suggest additional lines of research, including extensions to the multi-way (tensor) context using the framework established by Allen [26] or to other widely-used multivariate analysis techniques, such as partial least squares (PLS), canonical correlation analysis (CCA), and linear discriminant analysis (LDA).

7. REFERENCES

[1] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009. DOI: 10.1198/jasa.2009.0121.

[2] B. W. Silverman, "Smoothed functional principal components analysis by choice of norm," *Annals of Statistics*, vol. 24, no. 1, pp. 1–24, 1996. DOI: 10.1214/aos/1033066196.

[3] J. Z. Huang, H. Shen, and A. Buja, "Functional principal components analysis via penalized rank one approximation," *Electronic Journal of Statistics*, vol. 2, pp. 678–695, 2008. DOI: 10.1214/08-EJS218.

[4] ——, "The analysis of two-way functional data using two-way regularized singular value decompositions," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1609–1620, 2009. DOI: 10.1198/jasa.2009.tm08024.

[5] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009. DOI: 10.1093/biostatistics/kxp008.

[6] G. I. Allen and M. Maletić-Savatić, "Sparse non-negative generalized PCA with applications to metabolomics," *Bioinformatics*, vol. 27, no. 21, pp. 3029–3035, 2011. DOI: 10.1093/bioinformatics/btr522.

[7] G. I. Allen, L. Grosenick, and J. Taylor, "A generalized least-square matrix decomposition," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 145–159, 2014. DOI: 10.1080/01621459.2013.852978.

[8] P. Hall, "Principal component analysis for functional data: Methodology, theory, and discussion," in *The Oxford Handbook of Functional Data Analysis*, F. Ferraty and Y. Romain, Eds., 1st, Oxford University Press, 2011, pp. 210–235, ISBN: 978-0-199-56844-4. DOI: 10.1093/oxfordhb/9780199568444.013.8.

[9] H. Zou and L. Xue, "A selective overview of sparse principal component analysis," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1311–1320, 2018. DOI: 10.1109/JPROC.2018.2846588.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[11] 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*, vol. 67, no. 1, pp. 91–108, 2005. DOI: 10.1111/j.1467-9868.2005.00490.x.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006. DOI: 10.1111/j.1467-9868.2005.00532.x.

[13] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011. DOI: 10.1214/11-AOS878.

[14] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès, "SLOPE – adaptive variable selection via convex optimization," *Annals of Applied Statistics*, vol. 9, no. 3, pp. 1103–1140, 2015. DOI: 10.1214/15-AOAS842.

[15] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1015–1034, 2008. DOI: 10.1016/j.jmva.2007.06.007.

[16] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009. DOI: 10.1137/080716542.

[17] L. Mackey, "Deflation methods for sparse PCA," in *NIPS 2008: Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2008, pp. 1017–1024. Available at: https://papers.nips.cc/paper/3575-deflation-methods-for-sparse-pca.

[18] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011. DOI: 10.1137/090771806.

[19] K. Kato, "On the degrees of freedom in shrinkage estimation," *Journal of Multivariate Analysis*, vol. 100, no. 7, pp. 1338–1352, 2009. DOI: 10.1016/j.jmva.2008.12.002.

[20] R. J. Tibshirani and J. Taylor, "Degrees of freedom in lasso problems," *Annals of Statistics*, vol. 40, no. 2, pp. 1198–1232, 2012. DOI: 10.1214/12-AOS1003.

[21] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, "Biclustering via sparse singular value decomposition," *Biometrics*, vol. 66, no. 4, pp. 1087–1095, 2010. DOI: 10.1111/j.1541-0420.2010.01392.x.

[22] D. Dua and E. Karra Taniskidou, *UCI Machine Learning Repository*. Available at: http://archive.ics.uci.edu/ml.

[23] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," in *NIPS 1995: Advances in Neural Information Processing Systems 8*, D. S. Touretzsky, M. C. Mozer, and M. E. Hasselmo, Eds., 1995, pp. 145–151. Available at: https://papers.nips.cc/paper/1091-independent-component-analysis-of-electroencephalographic-data.

[24] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000. DOI: 10.1016/S0893-6080(00)00026-5.

[25] D. Shen, H. Shen, and J. S. Marron, "Consistency of sparse PCA in high dimension, low sample size contexts," *Journal of Multivariate Analysis*, vol. 115, pp. 317–333, 2013. DOI: 10.1016/j.jmva.2012.10.007.

[26] G. I. Allen, "Sparse higher-order principal components analysis," in *AISTATS 2012: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, vol. 22, Canary Islands, Spain: PMLR, 2012, pp. 27–36. Available at: http://proceedings.mlr.press/v22/allen12.html.

Supplementary Materials

A. PROOFS

Before proving the major results stated in the main body of the paper, we give three lemmas:

Lemma 1. *Suppose $f(\boldsymbol{x}) : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ is a non-negative function and is positive homogeneous of order one,* i.e., *$f(c\boldsymbol{x}) = cf(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^p$ and all $c \geq 0$. Then, if $\tilde{\nabla} f(\boldsymbol{x})$ is a sub-gradient of f at $\boldsymbol{x}$, then $\tilde{\nabla} f(c\boldsymbol{x})$ is also a sub-gradient for all $c \geq 0$.*

Proof. This follows immediately from the definition of a sub-gradient and the assumption of positive homogeneity. If $\tilde{\nabla} f(\boldsymbol{x})$ is a sub-gradient of f and $\boldsymbol{x}$, then we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \tilde{\nabla} f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \quad \forall y \in \mathbb{R}^p$$

Substitute $\boldsymbol{x} \to c\boldsymbol{x}$ and $\boldsymbol{y} \to c\boldsymbol{y}$ for arbitrary $c \geq 0$ to obtain

$$f(c\boldsymbol{y}) \geq f(c\boldsymbol{x}) + \tilde{\nabla} f(c\boldsymbol{x})^T (c\boldsymbol{y} - c\boldsymbol{x}) \quad \forall y \in \mathbb{R}^p.$$

Direct simplification yields

$$\begin{aligned} f(c\boldsymbol{y}) &\geq f(c\boldsymbol{x}) + \tilde{\nabla} f(c\boldsymbol{x})^T (c\boldsymbol{y} - c\boldsymbol{x}) \quad \forall \boldsymbol{y} \in \mathbb{R}^p \\ cf(\boldsymbol{y}) &\geq cf(\boldsymbol{x}) + c\tilde{\nabla} f(c\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \quad \forall \boldsymbol{y} \in \mathbb{R}^p \\ f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \tilde{\nabla} f(c\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \quad \forall \boldsymbol{y} \in \mathbb{R}^p \end{aligned}$$

which implies that $\tilde{\nabla} f(c\boldsymbol{x})$ is also a sub-gradient of f and $\boldsymbol{x}$. □

Lemma 2. *Suppose $(\boldsymbol{u}, \boldsymbol{v})$ are a global maximum of the SFPCA Problem (3). Then $(\boldsymbol{u}, \boldsymbol{v})$ satisfy the following Karush-Kuhn-Tucker (KKT) conditions:*

$$\begin{aligned} \boldsymbol{X}\boldsymbol{v} - \lambda_{\boldsymbol{u}} \tilde{\nabla}_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) - 2\gamma_{\boldsymbol{u}} \boldsymbol{S}_{\boldsymbol{u}} \boldsymbol{u} &= 0 \quad (\boldsymbol{u}\text{-stationarity}) \\ \gamma_{\boldsymbol{u}}(\|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} - 1) &= 0 \quad (\boldsymbol{u}\text{-complementary slackness}) \\ \boldsymbol{u}^T \boldsymbol{X} - \lambda_{\boldsymbol{v}} \tilde{\nabla}_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v}) - 2\gamma_{\boldsymbol{v}} \boldsymbol{S}_{\boldsymbol{v}} \boldsymbol{v} &= 0 \quad (\boldsymbol{v}\text{-stationarity}) \\ \gamma_{\boldsymbol{v}}(\|\boldsymbol{v}\|_{\boldsymbol{S}_{\boldsymbol{v}}} - 1) &= 0 \quad (\boldsymbol{v}\text{-complementary slackness}) \\ \gamma_{\boldsymbol{u}} &\geq 0 \quad (\boldsymbol{u}\text{-dual feasibility}) \\ \gamma_{\boldsymbol{v}} &\geq 0 \quad (\boldsymbol{v}\text{-dual feasibility}) \\ \|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} &\leq 1 \quad (\boldsymbol{u}\text{-primal feasibility}) \\ \|\boldsymbol{v}\|_{\boldsymbol{S}_{\boldsymbol{u}}} &\leq 1 \quad (\boldsymbol{v}\text{-primal feasibility}) \end{aligned}$$

where $\gamma_{\boldsymbol{u}}$ and $\gamma_{\boldsymbol{v}}$ are the dual variables associated with the inequality constraints of the SFPCA Problem (3) and $\tilde{\nabla} f(\boldsymbol{x})$ denotes an arbitrary sub-gradient of f and $\boldsymbol{x}$: that is, any value satisfying $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \tilde{\nabla} f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{y} \in \mathbb{R}$.

Proof. Despite the non-convexity of the SFPCA Problem (3), many of the classical results of convex analysis, including the KKT conditions, can be established for local minima under additional assumptions. Chapter 5 of Bertsekas *et al.* [1] gives an elegant presentation of these results. In particular, we note that the SFPCA Problem (3) satisfies their CQ5c, a variant of Slater's condition [2], for any local maximum as the point $(\boldsymbol{0}, \boldsymbol{0})$ is clearly strictly feasible. Additionally, we note that the feasible set $\overline{\mathbb{B}}^n_{\boldsymbol{S}_{\boldsymbol{u}}} \times \overline{\mathbb{B}}^p_{\boldsymbol{S}_{\boldsymbol{v}}}$ is clearly *regular* in their sense of having well-behaved normal and (polar) tangent cones (see, *e.g.*, their Definition 4.6.3). Since any global optimum must be a local optimum, the desired result follows. (The top right portion of their Figure 5.5.2 of Bertsekas *et al.* [1] is useful in following their presentation.) □

Lemma 3. *Suppose $P_{\boldsymbol{u}} : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ is a positive-homogeneous function of order one. Let*

$$\boldsymbol{u}^* = \begin{cases} \hat{\boldsymbol{u}}/\|\hat{\boldsymbol{u}}\|_{\boldsymbol{S}_{\boldsymbol{u}}} & \text{where } \|\boldsymbol{u}\|_{\boldsymbol{S}_{\boldsymbol{u}}} > 0 \\ \boldsymbol{0} & \text{otherwise} \end{cases} \quad \text{where } \hat{\boldsymbol{u}} \text{ is a stationary point of} \quad \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{v} - \boldsymbol{u}\|_2^2 + \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) + \frac{\alpha}{2}\boldsymbol{u}^T \boldsymbol{\Omega}_{\boldsymbol{u}} \boldsymbol{u} \tag{5}$$

Then $\boldsymbol{u}^*$ is a stationary point of

$$\operatorname*{arg\,max}_{\boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S}_{\boldsymbol{u}}}} \boldsymbol{u}^T \boldsymbol{X}\boldsymbol{v} - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) \tag{6}$$

(Note that Problem (5) is Problem (4) from the main text restated here for convenience.) Additionally, if $P_{\boldsymbol{u}}$ is convex, then $\hat{\boldsymbol{u}}$ and $\boldsymbol{u}^*$ are global optima of their respective subproblems.

Proof. We note that this proof follows the proof of Theorem 2 of Allen *et al.* [3]. Following the proof of Lemma 2 holding $\boldsymbol{v}$ fixed (and feasible), we have the following KKT conditions for Problem (6):

$$\begin{aligned} 2\gamma_{\boldsymbol{u}}\boldsymbol{S_u}\boldsymbol{u}^* - \boldsymbol{X}\boldsymbol{v} + \lambda_{\boldsymbol{u}}\tilde{\nabla}P_{\boldsymbol{u}}(\boldsymbol{u}^*) = 0 \quad &\text{(stationarity)} \\ \gamma_{\boldsymbol{u}}(\|\boldsymbol{u}^*\|_{\boldsymbol{S_u}} - 1) = 0 \quad &\text{(complementary slackness).} \end{aligned}$$

Similarly, the KKT conditions for $\hat{\boldsymbol{u}}$ in Problem (5) yield:

$$0 = -(\boldsymbol{X}\boldsymbol{v} - \hat{\boldsymbol{u}}) + \lambda_{\boldsymbol{u}}\tilde{\nabla}P_{\boldsymbol{u}}(\hat{\boldsymbol{u}}) + \alpha\boldsymbol{\Omega_u}\hat{\boldsymbol{u}} \implies \boldsymbol{S_u}\hat{\boldsymbol{u}} - \boldsymbol{X}\boldsymbol{v} + \lambda_{\boldsymbol{u}}\tilde{\nabla}P_{\boldsymbol{u}}(\hat{\boldsymbol{u}}) = 0$$

where $\hat{\boldsymbol{u}} + \alpha\Omega\hat{\boldsymbol{u}} = \boldsymbol{S_u}\hat{\boldsymbol{u}}$. Comparing the stationarity conditions for $\boldsymbol{u}^*$ and $\hat{\boldsymbol{u}}$, we see that they are equivalent up to the $2\gamma_{\boldsymbol{u}}$ term.

Let $\tilde{\boldsymbol{u}} = \hat{\boldsymbol{u}}/2\gamma_{\boldsymbol{u}}$. Then the KKT conditions of Problem (5) imply:

$$\begin{aligned} 0 &= \boldsymbol{S_u}\hat{\boldsymbol{u}} - \boldsymbol{X}\boldsymbol{v} + \lambda_{\boldsymbol{u}}\tilde{\nabla}P_{\boldsymbol{u}}(\hat{\boldsymbol{u}}) \\ &= 2\gamma_{\boldsymbol{u}}\boldsymbol{S_u}\tilde{\boldsymbol{u}} - \boldsymbol{X}\boldsymbol{v} + \lambda_{\boldsymbol{u}}\tilde{\nabla}P_{\boldsymbol{u}}(2\gamma_{\boldsymbol{u}}\tilde{\boldsymbol{u}}) \\ &= 2\gamma_{\boldsymbol{u}}\boldsymbol{S_u}\tilde{\boldsymbol{u}} - \boldsymbol{X}\boldsymbol{v} + \lambda_{\boldsymbol{u}}\tilde{\nabla}P_{\boldsymbol{u}}(\tilde{\boldsymbol{u}}) \end{aligned}$$

where the constant of $2\gamma_{\boldsymbol{u}}$ appearing in the sub-gradient could be removed using Lemma 1. From this we see that $\tilde{\boldsymbol{u}}$ satisfies the stationarity conditions for Problem (6). Hence, if we take $(\boldsymbol{u}^*, \gamma_{\boldsymbol{u}}) = (\hat{\boldsymbol{u}}/\|\hat{\boldsymbol{u}}\|_{\boldsymbol{S_u}}, \|\hat{\boldsymbol{u}}\|_{\boldsymbol{S_u}}/2)$, we have a solution to the KKT conditions for Problem (6), implying that we have a local solution. Additionally, if $P_{\boldsymbol{u}}(\cdot)$ is convex, then Problem (6) is concave, so the KKT conditions imply global optimality.

More intuitively, if we compare the stationarity conditions for $\boldsymbol{u}^*$ and $\hat{\boldsymbol{u}}$ directly, we see that they differ only by the leading constant factor of $2\gamma_{\boldsymbol{u}}$, suggesting that $\boldsymbol{u}^* \propto \hat{\boldsymbol{u}}$. Since we know $\boldsymbol{u}^*$ is a unit-vector under the $\boldsymbol{S_u}$-norm, we can guess $\boldsymbol{u}^* = \hat{\boldsymbol{u}}/\|\hat{\boldsymbol{u}}\|_{\boldsymbol{S_u}}$, which, when substituted into the KKT conditions, yields $\gamma_{\boldsymbol{u}} = \|\hat{\boldsymbol{u}}\|_{\boldsymbol{S_u}}$. □

With these results in hand, we are now ready to prove the main results of our paper, which we restate here for convenience.

Theorem 1. *Suppose Assumption 1 holds and let $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ be the optimal points of the SFPCA problem (3). Then the following hold:*

(i) There exist values $\lambda_{\boldsymbol{u}}^{\max}$ and $\lambda_{\boldsymbol{v}}^{\max}$ such that, if $\lambda_{\boldsymbol{u}} \geq \lambda_{\boldsymbol{u}}^{\max}$ or if $\lambda_{\boldsymbol{v}} \geq \lambda_{\boldsymbol{v}}^{\max}$, then the solution to Problem (3) is trivial in the sense $(\boldsymbol{u}^*, \boldsymbol{v}^*) = (\boldsymbol{0}, \boldsymbol{0})$.

(ii) If $\lambda_{\boldsymbol{u}} < \lambda_{\boldsymbol{u}}^{\max}$ and $\lambda_{\boldsymbol{v}} < \lambda_{\boldsymbol{v}}^{\max}$, the SFPCA solution $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ depends on all (non-zero) regularization parameters.

(iii) $\|\boldsymbol{u}^*\|_{\boldsymbol{S_u}}$ is equal to either 1 or 0, with the latter occurring only when $\lambda_{\boldsymbol{u}} \geq \lambda_{\boldsymbol{u}}^{\max}$ or $\lambda_{\boldsymbol{v}} \geq \lambda_{\boldsymbol{v}}^{\max}$. An analogous result holds for $\boldsymbol{v}^*$.

(iv) $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ do not suffer from scale non-identifiability. (That is, $(c\boldsymbol{u}^*, c^{-1}\boldsymbol{v}^*)$ is not a solution for any $c \geq 0$ except $c = 1$.)

Proof of Theorem 1. Throughout the following, we continue the notation used in the proof of Lemma 3 and let $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ denote solutions to the SFPCA problem (3), while $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ denote solutions to Problem (5) and its analogue in $\boldsymbol{v}$:

$$\hat{\boldsymbol{v}} = \underset{\boldsymbol{v}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{2}\|\boldsymbol{u}^T\boldsymbol{X} - \boldsymbol{v}\|_2^2 + \lambda_{\boldsymbol{v}}P_{\boldsymbol{v}}(\boldsymbol{v}) + \frac{\alpha}{2}\boldsymbol{v}\boldsymbol{\Omega_v}\boldsymbol{v}$$

Part (i) From Lemma 3, we have that $\boldsymbol{u}^* = 0$ if and only if $\hat{\boldsymbol{u}} = 0$. The KKT conditions for Problem (5) show this occurs only when

$$\lambda_{\boldsymbol{u}}\tilde{\nabla}P_{\boldsymbol{u}}(\boldsymbol{0}) = \boldsymbol{X}\boldsymbol{v}.$$

Hence, for any fixed $\boldsymbol{v}$, we can find a value $\lambda_{\boldsymbol{u}}^{\boldsymbol{v},\max}$, such that $\lambda_{\boldsymbol{u}} \geq \lambda_{\boldsymbol{u}}^{\boldsymbol{v},\max}$ yields an all-zero solution. Taking the maximum over all $\boldsymbol{v} \in \overline{\mathbb{B}}_{\boldsymbol{S_v}}^p$, we obtain $\lambda_{\boldsymbol{u}}^{\max}$ as desired. An analogous result holds for $\lambda_{\boldsymbol{v}}^{\max}$.

Additionally, we note that if $\boldsymbol{u} = \boldsymbol{0}$, then $\boldsymbol{v} = \boldsymbol{0}$ satisfies the KKT conditions given in Lemma 2 and hence is a solution. Putting these together, we note that if $\lambda_{\boldsymbol{u}} \geq \lambda_{\boldsymbol{u}}^{\max}$ or if $\lambda_{\boldsymbol{v}} \geq \lambda_{\boldsymbol{v}}^{\max}$, then $(\boldsymbol{u}^*, \boldsymbol{v}^*) = (\boldsymbol{0}, \boldsymbol{0})$, as desired.

Part (ii) Now, we assume $\lambda_{\boldsymbol{u}} < \lambda_{\boldsymbol{u}}^{\max}$ and $\lambda_{\boldsymbol{v}} < \lambda_{\boldsymbol{v}}^{\max}$, so $\boldsymbol{u}^* \neq \boldsymbol{0}$ and $\boldsymbol{v}^* \neq 0$, and $\alpha_{\boldsymbol{u}}, \alpha_{\boldsymbol{v}} > 0$. By the $\boldsymbol{u}$-stationary term of the KKT conditions given in Lemma 2, it is clear that $\boldsymbol{u}^*$ depends on both $\lambda_{\boldsymbol{u}}$ and $\alpha_{\boldsymbol{u}}$, by way of $\boldsymbol{S_u}$, as well as $\boldsymbol{v}^*$. A similar argument shows that $\boldsymbol{v}^*$ depends on both $\lambda_{\boldsymbol{v}}$ and $\alpha_{\boldsymbol{v}}$ as well as $\boldsymbol{u}^*$, so transitively both $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ depend on all (non-zero) regularization parameters.

Part (iii) Consider the $\boldsymbol{u}$-complimentary slackness condition given in Lemma 2, which implies that $\gamma_{\boldsymbol{u}} > 0$ if and only if $\|\boldsymbol{u}^*\|_{\boldsymbol{S_u}} = 1$. In the proof of Lemma 3, we showed that solutions to the SFPCA KKT conditions are of the form $(\boldsymbol{u}^*, \gamma_{\boldsymbol{u}}) = (\hat{\boldsymbol{u}}/\|\hat{\boldsymbol{u}}\|_{\boldsymbol{S_u}}, \|\hat{\boldsymbol{u}}\|_{\boldsymbol{S_u}}/2)$. Hence, $\gamma_{\boldsymbol{u}} = 0$ if and only if $\hat{\boldsymbol{u}} = \boldsymbol{0}$, which, by Part (i), occurs when $\lambda_{\boldsymbol{u}} \geq \lambda_{\boldsymbol{u}}^{\max}$ or $\lambda_{\boldsymbol{v}} \geq \lambda_{\boldsymbol{v}}^{\max}$. Putting this together, if the $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ are non-zero, then the boundary conditions must hold with $\|\boldsymbol{u}^*\|_{\boldsymbol{S_u}} = \|\boldsymbol{v}^*\|_{\boldsymbol{S_v}} = 1$.

Part (iv) As shown in Part (iii), for non-trivial solutions we have $\|\boldsymbol{u}^*\|_{\boldsymbol{S_u}} = \|\boldsymbol{v}^*\|_{\boldsymbol{S_v}} = 1$, so the SFPCA problem does not suffer from scale non-identifiability: that is, if $(\boldsymbol{u}, \boldsymbol{v})$ is a solution, we do not have additional solutions of the form $(c\boldsymbol{u}, c^{-1}\boldsymbol{v})$ for $c > 0$. If $P_{\boldsymbol{u}}, P_{\boldsymbol{v}}$ are even functions (that is $P_{\boldsymbol{u}}(-\boldsymbol{u}) = P_{\boldsymbol{u}}(\boldsymbol{u})$ and $P_{\boldsymbol{v}}(-\boldsymbol{v}) = P_{\boldsymbol{v}}(\boldsymbol{v})$ for all $\boldsymbol{u}, \boldsymbol{v}$), the SFPCA problem still has a sign non-identifiability.

□

Theorem 2. *Suppose Assumption 1 holds and let $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ be the optimal points of the SFPCA problem (3). Then the following hold (up to a sign factor and unit scaling):*

(i) If $\lambda_{\boldsymbol{u}}, \lambda_{\boldsymbol{v}}, \alpha_{\boldsymbol{u}}, \alpha_{\boldsymbol{v}} = 0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are the first left and right singular vectors of $\boldsymbol{X}$.

(ii) If $\lambda_{\boldsymbol{u}}, \alpha_{\boldsymbol{u}}, \alpha_{\boldsymbol{v}} = 0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the SPCA solution of Shen and Huang [4].

(iii) If $\alpha_{\boldsymbol{u}}, \alpha_{\boldsymbol{v}} = 0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the two-way SPCA solution in Allen et al. *[3], itself a special case of two-way sparse GPCA with the generalizing operators $\boldsymbol{Q}, \boldsymbol{R}$ both identity matrices. (This is also the Lagrangian form of Witten* et al. *[5].)*

(iv) If $\lambda_{\boldsymbol{u}}, \lambda_{\boldsymbol{v}}, \alpha_{\boldsymbol{u}} = 0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the FPCA solution of Silverman [6] and Huang et al. *[7].*

(v) If $\lambda_{\boldsymbol{u}}, \lambda_{\boldsymbol{v}} = 0$, then $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ are equivalent to the two-way FPCA solution of Huang et al. *[8].*

For parts (ii) and (iii), equivalencies hold for the appropriate $P_{\boldsymbol{u}}(\cdot)$ and $P_{\boldsymbol{v}}(\cdot)$ employed in the referenced papers.

Proof of Theorem 2. We establish the equivalence of several cases of SFPCA with approaches previously proposed in the literature.

Part (i) $\lambda_{\boldsymbol{u}} = \lambda_{\boldsymbol{v}} = \alpha_{\boldsymbol{u}} = \alpha_{\boldsymbol{v}} = 0$. In this case, the SFPCA Problem (3) simplifies to

$$\underset{\substack{\boldsymbol{u}\in\mathbb{R}^n:\boldsymbol{u}^T\boldsymbol{u}\leq 1\\ \boldsymbol{v}\in\mathbb{R}^p:\boldsymbol{v}^T\boldsymbol{v}\leq 1}}{\arg\max} \quad \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v}$$

which we recognize as the Singular Value Problem (1), which defines standard PCA.

Part (ii) $\lambda_{\boldsymbol{u}} = \alpha_{\boldsymbol{u}} = \alpha_{\boldsymbol{v}} = 0$. The SPCA estimator of Shen and Huang [4] is given by

$$\boldsymbol{v}^* = \hat{\boldsymbol{v}}/\|\hat{\boldsymbol{v}}\| \text{ where } \hat{\boldsymbol{u}}, \hat{\boldsymbol{v}} = \underset{\substack{\boldsymbol{u}\in\mathbb{B}^n\\ \boldsymbol{v}\in\mathbb{R}^p}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}^T\|_F^2 + \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v})$$

Taking the KKT conditions with respect to $\boldsymbol{v}$, we obtain:

$$\boldsymbol{X}^T\hat{\boldsymbol{u}} - \hat{\boldsymbol{v}} - \lambda_{\boldsymbol{v}}\tilde{\nabla}P_{\boldsymbol{v}}(\hat{\boldsymbol{v}}) = 0$$

Comparing this to the KKT conditions for SFPCA derived in Lemma 2 with $\lambda_{\boldsymbol{u}} = \alpha_{\boldsymbol{u}} = \alpha_{\boldsymbol{v}} = 0$,

$$\begin{aligned}\boldsymbol{X}^T\hat{\boldsymbol{u}} - 2\gamma_{\boldsymbol{v}}\hat{\boldsymbol{v}} - \lambda_{\boldsymbol{v}}\tilde{\nabla}P_{\boldsymbol{v}}(\hat{\boldsymbol{v}}) &= 0\\ \gamma_{\boldsymbol{v}}(\|\hat{\boldsymbol{v}}\| - 1) &= 0,\end{aligned}$$

we see that the only difference is the factor of $2\gamma_{\boldsymbol{v}}$ in the stationarity conditions. As before, we define $\hat{\boldsymbol{v}} = 2\gamma_{\boldsymbol{v}}\tilde{\boldsymbol{v}}$ and re-write the SPCA stationarity conditions as

$$0 = \boldsymbol{X}^T\hat{\boldsymbol{u}} - \hat{\boldsymbol{v}} - \lambda_{\boldsymbol{v}}\tilde{\nabla}P_{\boldsymbol{v}}(\hat{\boldsymbol{v}}) = \boldsymbol{X}^T\hat{\boldsymbol{u}} - 2\gamma_{\boldsymbol{v}}\tilde{\boldsymbol{v}} - \lambda_{\boldsymbol{v}}\tilde{\nabla}P_{\boldsymbol{v}}(2\gamma\boldsymbol{v}\tilde{\boldsymbol{v}}) = \boldsymbol{X}^T\hat{\boldsymbol{u}} - 2\gamma_{\boldsymbol{v}}\tilde{\boldsymbol{v}} - \lambda_{\boldsymbol{v}}\tilde{\nabla}P_{\boldsymbol{v}}(\tilde{\boldsymbol{v}}),$$

where the final equality follows from Lemma 1. This clearly matches the $\boldsymbol{v}$-stationarity condition for SFPCA and the scaling step implies the complementary slackness condition holds, showing the two solutions are equivalent.

Part (iii) $\alpha_{\boldsymbol{u}} = \alpha_{\boldsymbol{v}} = 0$. In this case, the SFPCA Problem (3) simplifies to

$$\underset{\substack{\boldsymbol{u}\in\mathbb{R}^n:\boldsymbol{u}^T\boldsymbol{u}\leq 1\\ \boldsymbol{v}\in\mathbb{R}^p:\boldsymbol{v}^T\boldsymbol{v}\leq 1}}{\arg\max} \quad \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} - \lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}}P_{\boldsymbol{v}}(\boldsymbol{v})$$

which is clearly equivalent to the sparse GPCA method of Allen *et al.* [3, Equation 6] with the generalizing operators $\boldsymbol{Q}, \boldsymbol{R}$ both set equal to identity matrices. For non-convex problems such as SFPCA (3), it is not always the case that constraints can be re-written as Lagrange multipliers and penalty functions; conditions under which this is possible are discussed in Chapter 5 of Bertsekas *et al.* [1] and do indeed apply here. (See also the discussion in the proof of Lemma 2.)

Part (iv) $\lambda_{\boldsymbol{u}} = \lambda_{\boldsymbol{v}} = \alpha_{\boldsymbol{u}} = 0$. Huang *et al.* [7] consider a penalized regression formulation of FPCA:

$$\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}^T\|_F^2 + \alpha\boldsymbol{u}^T\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{\Omega}_{\boldsymbol{v}}\boldsymbol{v}.$$

They show that $\boldsymbol{v}$ of this formulation is equivalent to (a discretization of) the earlier FPCA formulation of Silverman [6]:

$$\underset{\boldsymbol{v}\in\mathbb{R}^p}{\arg\max}\, \boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to} \quad \boldsymbol{v}^T\boldsymbol{S}_{\boldsymbol{v}}\boldsymbol{v} = 1$$

We compare this to our SFPCA formulation with only $\alpha_{\boldsymbol{v}}$ non-zero:

$$\underset{\boldsymbol{u},\boldsymbol{v}}{\arg\max}\, \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} \quad \text{subject to} \quad \boldsymbol{u}^T\boldsymbol{u} \leq 1 \text{ and } \boldsymbol{v}^T\boldsymbol{S}_{\boldsymbol{v}}\boldsymbol{v} \leq 1$$

Examination of the KKT conditions reveals that, for given $\boldsymbol{v}$, the above criterion is maximized by taking $\boldsymbol{u} = \boldsymbol{X}\boldsymbol{v}/\sqrt{\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v}}$. Substituting this into the above, we see that SFPCA simplifies to

$$\arg\max_{\boldsymbol{v}} \sqrt{\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v}} \quad \text{subject to} \quad \boldsymbol{v}^T\boldsymbol{S}_{\boldsymbol{v}}\boldsymbol{v} \leq 1$$

As shown in Theorem 1, the constraint must hold tightly (since there are no sparsity penalties), and the $\sqrt{\cdot}$ transform is monotonic, so this is clearly equivalent to the FPCA formulation of Silverman [6], which establishes the desired equivalence for the right singular vectors. For the left singular vectors, Huang *et al.* [7] show that the solution to their FPCA formulation is obtained by an iterative method containing the update $\boldsymbol{u} := \boldsymbol{X}\boldsymbol{v}/\|\boldsymbol{v}\|_{\boldsymbol{S}_{\boldsymbol{v}}}$; this is exactly the same as our expression for $\boldsymbol{u}$ *modulo* a normalizing factor.

Part (v) $\lambda_{\boldsymbol{u}} = \lambda_{\boldsymbol{v}} = 0$. Huang *et al.* [8] consider two-way FPCA as:

$$\arg\max_{\boldsymbol{u}\in\mathbb{R}^n, \boldsymbol{v}\in\mathbb{R}^p} \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} - \frac{\boldsymbol{u}^T(\boldsymbol{I}+\boldsymbol{\Omega}_{\boldsymbol{u}})\boldsymbol{u} \cdot \boldsymbol{v}^T(\boldsymbol{I}+\boldsymbol{\Omega}_{\boldsymbol{v}})\boldsymbol{v}}{2}$$

This gives stationarity conditions of the form $\boldsymbol{u} \propto (\boldsymbol{I}+\boldsymbol{\Omega}_{\boldsymbol{u}})^{-1}\boldsymbol{X}\boldsymbol{v}$ and $\boldsymbol{v} \propto (\boldsymbol{I}+\boldsymbol{\Omega}_{\boldsymbol{v}})^{-1}\boldsymbol{X}^T\boldsymbol{v}$. (Note that Huang *et al.* [8] define their smoothing matrices $\boldsymbol{S}_{\boldsymbol{u}}, \boldsymbol{S}_{\boldsymbol{v}}$ as the multiplicative inverses of the definitions we use.) With $\lambda_{\boldsymbol{u}} = \lambda_{\boldsymbol{v}} = 0$, the SFPCA KKT conditions derived in Lemma 2 simplify to:

$$\begin{aligned}
\boldsymbol{X}\boldsymbol{v}^* - 2\gamma_{\boldsymbol{u}}\boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}^* &= 0 \\
\gamma_{\boldsymbol{u}}(\|\boldsymbol{u}^*\|_{\boldsymbol{S}_{\boldsymbol{u}}} - 1) &= 0 \\
\boldsymbol{X}^T\boldsymbol{u}^* - 2\gamma_{\boldsymbol{v}}\boldsymbol{S}_{\boldsymbol{v}}\boldsymbol{v}^* &= 0 \\
\gamma_{\boldsymbol{v}}(\|\boldsymbol{v}^*\|_{\boldsymbol{S}_{\boldsymbol{v}}} - 1) &= 0
\end{aligned}$$

From these, we find $\boldsymbol{u}^* \propto \boldsymbol{S}_{\boldsymbol{u}}^{-1}\boldsymbol{X}\boldsymbol{v}^*$ and $\boldsymbol{v}^* \propto \boldsymbol{S}_{\boldsymbol{v}}^{-1}\boldsymbol{X}^T\boldsymbol{u}^*$, which clearly match the two-way FPCA stationary conditions if we take $\alpha_{\boldsymbol{u}} = \alpha_{\boldsymbol{v}} = 1$, thereby establishing the desired equivalence. We note, however, that the scaling factors used by SFPCA and the method of Huang *et al.* [8] are different, as we take $\boldsymbol{u} = \boldsymbol{S}_{\boldsymbol{u}}^{-1}\boldsymbol{X}\boldsymbol{v}/\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{S}_{\boldsymbol{u}}^{-1}\boldsymbol{X}\boldsymbol{v}$ while they take $\boldsymbol{u} = \boldsymbol{S}_{\boldsymbol{u}}^{-1}\boldsymbol{X}\boldsymbol{v}/\boldsymbol{v}^T\boldsymbol{S}_{\boldsymbol{v}}\boldsymbol{v}$ and similarly for the $\boldsymbol{v}$-normalization. This change in scaling is essentially cosmetic, as it does not effect the direction or relative weights of the estimated principal components. □

Theorem 3. *Under Assumption 1, Algorithm 1 has the following properties:*

(i) Step 2(a) converges to a stationary point of

$$\arg\min_{\boldsymbol{u}\in\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{v}-\boldsymbol{u}\|_2^2 + \lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u}) + \frac{\alpha_{\boldsymbol{u}}}{2}\boldsymbol{u}^T\boldsymbol{\Omega}_{\boldsymbol{u}}\boldsymbol{u}. \tag{4}$$

Furthermore, if $P_{\boldsymbol{u}}$ is convex, the convergence is monotone, at an $\mathcal{O}(1/K)$ rate, and to a global solution. Step 2(b) converges analogously for $\boldsymbol{v}$ and $P_{\boldsymbol{v}}$.

(ii) If $P_{\boldsymbol{u}}$ is convex, Step 2(a) yields a global solution to (3), considering $\hat{\boldsymbol{v}}$ fixed; if $P_{\boldsymbol{u}}$ is non-convex, Step 2(a) yields a stationary point for $P_{\boldsymbol{u}}$, considering $\hat{\boldsymbol{v}}$ fixed. An analogous result holds for $\hat{\boldsymbol{v}}$ returned by Step 2(b), with $\hat{\boldsymbol{u}}$ considered fixed.

(iii) If $P_{\boldsymbol{u}}, P_{\boldsymbol{v}}$ are both convex, then $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}})$ returned by the SFPCA Algorithm (1) is both a coordinate-wise global maximum (Nash point) and a stationary point of Problem (3).

Proof of Theorem 3. We first note that the $\boldsymbol{u}$-subproblem (4) can be re-written as

$$\arg\min_{\boldsymbol{u}\in\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n} \frac{\boldsymbol{u}^T\boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}}{2} - \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v} + \lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u}) = \arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \underbrace{\frac{\boldsymbol{u}^T\boldsymbol{S}_{\boldsymbol{u}}\boldsymbol{u}}{2} - \boldsymbol{u}^T\boldsymbol{X}\boldsymbol{v}}_{\text{smooth}} + \underbrace{\lambda_{\boldsymbol{u}}P_{\boldsymbol{u}}(\boldsymbol{u}) + \iota_{\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n}(\boldsymbol{u})}_{\text{non-differentiable}}$$

where ι represents the (infinite) indicator of the feasible set: that is, $\iota_{\mathcal{X}}(x)$ is zero if x is an element of $\mathcal{X}$ and (positive) infinity otherwise. We note that the use of the indicator function here is justified despite possible non-convexity because it always holds as a tight constraint since have a feasible point at $\mathbf{0}$.

The first (smooth) term is strictly and strongly convex, since $\boldsymbol{S}_{\boldsymbol{u}} \succ 0$ by construction, and has a continuous gradient whose Lipschitz constant is given by the leading eigenvalue of $\boldsymbol{S}_{\boldsymbol{u}}$. We will make repeated use of the proximal mapping of the non-differentiable term, $\lambda_{\boldsymbol{u}}P_{\boldsymbol{u}} + \iota_{\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n}$, given by $\mathsf{prox}_f(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + f(\boldsymbol{z})$ for a given function f. For convex functions, the existence and uniqueness of the proximal mapping follow immediately from the properties of strongly convex functions; the properties of the proximal mapping were studied for a wide class of so-called *prox regular* non-convex functions by Poliquin and Rockafellar [9], [10]. Where the proximal mapping is not unique, any minimizer can be used in Algorithm 1. Gong *et al.* [11] give proximal operators for a range of widely-used convex and non-convex penalty functions.

We note a general result for any f satisfying the second part of Assumption 1 (positive-homogeneity):

$$\mathsf{prox}_{f+\iota_{\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + f(\boldsymbol{z}) + \iota_{\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n}(\boldsymbol{z}) = \arg\min_{\boldsymbol{z}\in\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + f(\boldsymbol{z}) = \mathsf{proj}_{\mathbb{B}_{\boldsymbol{S}_{\boldsymbol{u}}}^n}(\mathsf{prox}_f(\boldsymbol{x}))$$

where $\mathsf{proj}_{\mathcal{X}}(\boldsymbol{x})$ denotes the projection of $\boldsymbol{x}$ onto $\mathcal{X}$. This follows from Theorem 4 of Yu [12] where we take $h(\cdot) = \iota_{\mathbb{R}_{>1}}$ which is clearly an increasing function, and $\|\boldsymbol{x}\| = \boldsymbol{x}^T \boldsymbol{S}_{\boldsymbol{u}} \boldsymbol{x}$. (See also Corollary 1 of Yu [12].) Since we assume positive homogeneity of f, it is in the class of functions covered by that theorem and the desired result holds.

Part (i) Step 2(a) of Algorithm 1 is a standard proximal gradient iteration with fixed step-size applied to the $\boldsymbol{u}$-subproblem (4). If $P_{\boldsymbol{u}}$ is convex, then monotone $\mathcal{O}(1/K)$ convergence to a global solution follows from well-known results on proximal gradient methods: see, *e.g.*, Theorems 10.21 ($\mathcal{O}(1/K)$ convergence), 10.23 (Fejér Monotonicity), and 10.24 (convergence to a global optimum) of Beck [13]. If $P_{\boldsymbol{u}}$ is non-convex, convergence to a stationary point follows from Theorem 10.15(d) of Beck [13]. Additionally, we note that, even in the nonconvex setting, step 2(a) monotonically decreases the objective function of the $\boldsymbol{u}$-subproblem (4) Beck [13, Theorem 10.15(a)].

Part (ii) This follows immediately from Lemma 3 and Part (i).

Part (iii) We note that Algorithm 1 can be considered a block-coordinate ascent algorithm for the SFPCA problem (3), where a proximal gradient scheme is used to solve each subproblem. While SFPCA (3) is non-concave, it is block bi-concave in $\boldsymbol{u}$ and $\boldsymbol{v}$, allowing for certain convergence results to be used. In particular, for $P_{\boldsymbol{u}}, P_{\boldsymbol{v}}$ both convex, we can use the results of Gorski *et al.* [14, Theorem 4.7] to establish convergence to a so-called Nash point (coordinate-wise optimum) satisfying

$$f(\boldsymbol{u}, \boldsymbol{v}^*) \leq f(\boldsymbol{u}^*, \boldsymbol{v}^*) \quad \text{for all } \boldsymbol{u} \in \overline{\mathbb{B}}^n_{\boldsymbol{S}_{\boldsymbol{u}}}$$
$$f(\boldsymbol{u}^*, \boldsymbol{v}) \leq f(\boldsymbol{u}^*, \boldsymbol{v}^*) \quad \text{for all } \boldsymbol{v} \in \overline{\mathbb{B}}^p_{\boldsymbol{S}_{\boldsymbol{v}}}$$

where $f(\boldsymbol{u}, \boldsymbol{v})$ is the SFPCA objective $f(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{v} - \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) - \lambda_{\boldsymbol{v}} P_{\boldsymbol{v}}(\boldsymbol{v})$. (Theorem 2.3 of Xu and Yin [15] generalizes this approach to approximate solutions of the subproblem, at the cost of requiring strong convexity.)

To show that the output of Algorithm 1 is also a stationary point, we use the regularity analysis of Tseng [16]: in particular, we note that the smooth part of our objective ($f_0(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T \boldsymbol{X} \boldsymbol{y}$) is (Gâteaux-)differentiable everywhere, so Tseng's assumption A1 holds.[2] This establishes regularity everywhere, including at each coordinate-wise maximum, which implies each Nash point is also a stationary point. □

We conjecture, but do not prove here, a generalization of the above: if $P_{\boldsymbol{u}}$ or $P_{\boldsymbol{v}}$ are non-convex, Algorithm 1 is still guaranteed to converge to a coordinate-wise local maximum (local Nash point). Xu and Yin [17] prove a related result, establishing convergence to a critical point, but where only a single gradient step is taken instead of fully solving the $\boldsymbol{u}$- and $\boldsymbol{v}$-subproblems as we do in Algorithm 1.

Additionally, we note that experimental evidence suggests neither positive-homogeneity nor convexity of $P_{\boldsymbol{u}}, P_{\boldsymbol{v}}$ are required for convergence of Algorithm 1, though we are unable to provide a full proof. Similar results have been previously demonstrated for coordinate descent schemes applied to related problems [18, Theorem 4] [19, Proposition 1] [17, Theorem 3.1], though they do not consider constraints and require a quadratic smooth term which we do not have here.

Finally, we note that in Step 2(a) of Algorithm 1, we self-normalize $\boldsymbol{u}$ under the $\boldsymbol{S}_{\boldsymbol{u}}$-norm in order to obtain $\hat{\boldsymbol{u}}$ at each step. Algorithmically, this can be considered a projected gradient scheme, where projection is required to ensure feasibility at each step. In the non-sparse case ($\lambda_{\boldsymbol{u}} = 0$), this update has the closed form $\hat{\boldsymbol{u}} = \boldsymbol{S}_{\boldsymbol{u}}^{-1} \boldsymbol{X} \boldsymbol{v} / \|\boldsymbol{S}_{\boldsymbol{u}}^{-1} \boldsymbol{X} \boldsymbol{v}\|_{\boldsymbol{S}_{\boldsymbol{u}}} = \boldsymbol{S}_{\boldsymbol{u}}^{-1} \boldsymbol{X} \boldsymbol{v} / \|\boldsymbol{X} \boldsymbol{v}\|_{\boldsymbol{S}_{\boldsymbol{u}}^{-1}}$, which is closely related to the updates in the two-way functional PCA algorithm of Huang *et al.* [8], but using a different normalization. As Huang *et al.* [8] discuss, this is equivalent to the more standard "half-smoothing" approach popularized by Silverman [6]. (As Allen [20] discusses, this equivalence does not extend straightforwardly to the higher-order array (tensor) context.)

Theorem 4. *Suppose $P_{\boldsymbol{u}}(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$. Then an unbiased estimate degrees of freedom of the $\boldsymbol{u}$-subproblem (4) is given by*

$$\widehat{\mathrm{df}}(\hat{\boldsymbol{u}}) = \mathrm{Tr}\left[\left(\boldsymbol{I}_{|\mathcal{A}|} + \alpha_{\boldsymbol{u}} \boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}}\right)^{-1}\right] \approx \mathrm{Tr}\left[\boldsymbol{I}_{|\mathcal{A}|} - \alpha_{\boldsymbol{u}} \boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}}\right]$$

where $\mathcal{A}$ denotes the indices of the estimated non-zero elements of $\hat{\boldsymbol{u}}$ and $\boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}}$ denotes the corresponding submatrix of $\boldsymbol{\Omega}_{\boldsymbol{u}}$. Hence, the approximate BIC to be optimized for subproblem (4) is given by

$$\mathrm{BIC}(\hat{\boldsymbol{u}}) = \log\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{v} - \hat{\boldsymbol{u}}\|_2^2\right] + \frac{1}{n}\log(n)\,\widehat{\mathrm{df}}(\hat{\boldsymbol{u}}).$$

Proof of Theorem 4. Consider the $\boldsymbol{u}$-update with ℓ_1-penalization [21]. In this case, the $\boldsymbol{u}$-subproblem (4) is essentially a *generalized* elastic net problem, [22] which can be analyzed using the techniques of Tibshirani and Taylor [23]. In particular, we re-write Problem (4) as a lasso problem with an augmented design matrix:

$$\arg\min_{\boldsymbol{u}} \frac{1}{2}\left\|\begin{pmatrix}\boldsymbol{X}\boldsymbol{v} \\ \boldsymbol{0}\end{pmatrix} - \begin{pmatrix}\boldsymbol{I} \\ \mathsf{chol}(\alpha_{\boldsymbol{u}}\boldsymbol{\Omega}_{\boldsymbol{u}})\end{pmatrix}\boldsymbol{u}\right\|_2^2 + \lambda_{\boldsymbol{u}}\|\boldsymbol{u}\|_1$$

where $\tilde{\boldsymbol{X}}$ is the augmented design matrix $\tilde{\boldsymbol{X}} = \begin{pmatrix}\boldsymbol{I} & \mathsf{chol}(\alpha_{\boldsymbol{u}}\Omega_{\boldsymbol{u}})^T\end{pmatrix}^T$. Then the degrees of freedom are given by

$$\mathrm{df} = \mathbb{E}\left[\mathrm{Tr}\left((\tilde{\boldsymbol{X}}_{\mathcal{A}}^T \tilde{\boldsymbol{X}}_{\mathcal{A}})^{-1}\right)\right].$$

[2] Tseng's treatment of constraints is somewhat unclear here, but we incorporate the unit ellipse constraints as indicator functions in the non-convex penalty portion of the problem, as discussed above, so dom $f_0 = \mathbb{R}^n \times \mathbb{R}^p$.

Note that the general form of their estimator is $\mathrm{Tr}(\boldsymbol{X}_{\mathcal{A}}(\tilde{\boldsymbol{X}}_{\mathcal{A}}^T\tilde{\boldsymbol{X}}_{\mathcal{A}})^{-1}\boldsymbol{X}_{\mathcal{A}})$ but we omit the outer terms as they are simply $\boldsymbol{I}$ for this problem. The sample value of this quantity gives an unbiased estimate of the degrees of freedom:

$$\widehat{\mathrm{df}} = \mathrm{Tr}\left((\tilde{\boldsymbol{X}}_{\mathcal{A}}^T\tilde{\boldsymbol{X}}_{\mathcal{A}})^{-1}\right) = \mathrm{Tr}\left((\boldsymbol{I}_{|\mathcal{A}|} + \alpha\boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}})^{-1}\right).$$

Rather than calculating the inverse, we substitute the first two terms of the Taylor expansion $(\boldsymbol{I}+\boldsymbol{A})^{-1} = I - \boldsymbol{A} + \boldsymbol{A}^2 - \boldsymbol{A}^3 + \dots$ to get

$$\widehat{\mathrm{df}} \approx \mathrm{Tr}\left(\boldsymbol{I}_{|\mathcal{A}|} - \alpha\boldsymbol{\Omega}_{\boldsymbol{u}}^{\mathcal{A}}\right).$$

The approximate BIC can then be obtained by substitution into the standard BIC formula [24], [25], using the maximum likelihood estimate of the residual variance $\hat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{X}\boldsymbol{v} - \hat{\boldsymbol{u}}\|_2^2 = \mathrm{RSS}/n$:

$$\begin{aligned}
\text{log-likelihood} &= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(u_i - \boldsymbol{x}_i^T\boldsymbol{v})^2}{2\sigma^2} \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\|\boldsymbol{u} - \boldsymbol{X}\boldsymbol{v}\|_2^2}{2\sigma^2} \\
\text{log-likelihood}|_{\hat{\sigma}^2=\mathrm{RSS}/n} &= -\frac{n}{2}\log(\mathrm{RSS}/n) - \frac{n}{2}\log(2\pi) - \frac{n}{2} \\
-2 * \text{log-likelihood}|_{\hat{\sigma}^2=\mathrm{RSS}/n} &= n\log(\mathrm{RSS}/n) + n\log(2\pi) + n \\
\implies \mathrm{BIC}(\hat{\boldsymbol{u}}) &= \log\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{v} - \hat{\boldsymbol{u}}\|_2^2\right] + \frac{1}{n}\log(n)\,\widehat{\mathrm{df}}(\hat{\boldsymbol{u}})
\end{aligned}$$

where the $n\log(2\pi)$ and n constant terms can be omitted in the BIC criterion. □

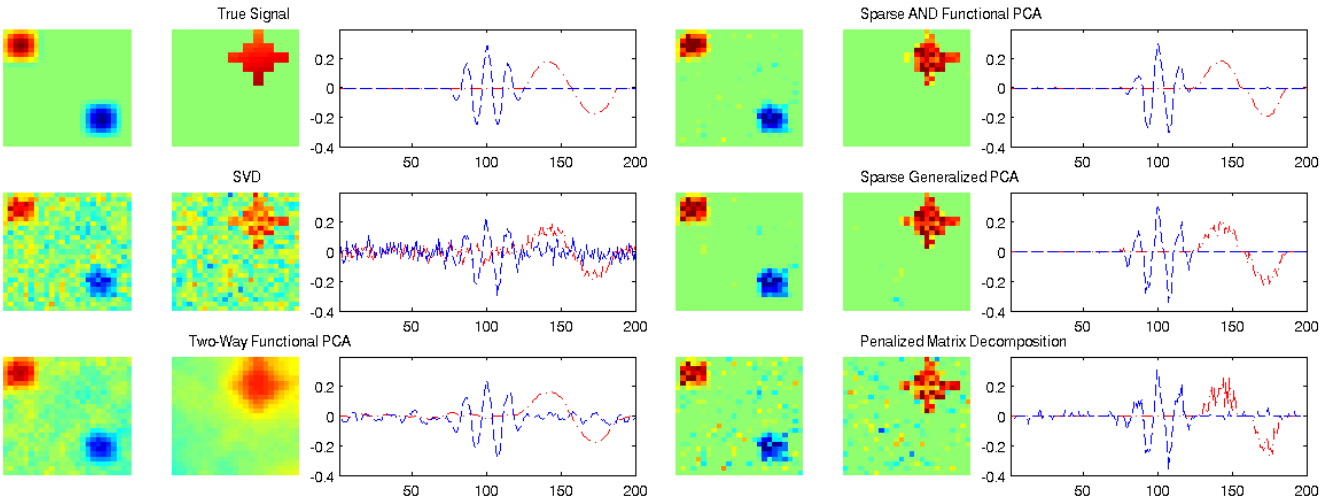

Fig. 4. Simulated factors used for the simulation study in Section B.1 and estimates thereof: for each sub-figure, the first two panels are the left singular vectors ($\boldsymbol{u}_1, \boldsymbol{u}_2$), while the third panel shows the right singular vectors, $\boldsymbol{v}_1$ (red, dotted-dashed) and $\boldsymbol{v}_2$ (blue, dashed). While SFPCA and SGPCA both perform well on the left singular vectors, only SFPCA is able to simultaneously identify the spatial sparsity and smooth structure of the sinusoidal pulses in the right singular vectors.

B. ADDITIONAL RESULTS

In this section, we extend the results given in Sections 4 and 5.

B.1. Two-Way Simulation Study

The simulation presented in Section 4 (Table 1 and Figure 2) contain exhibit smooth and sparse structure in the right singular vectors ($\boldsymbol{v}$), but not in the left singular vectors ($\boldsymbol{u}$), which are selected from the unit sphere randomly (Haar measure). In this section, we demonstrate the performance of SFPCA on data exhibiting smoothness and sparstity in both $\boldsymbol{u}$ and $\boldsymbol{v}$ in a rank-2 model.

The true factors in this simulation are inspired by neuroimaging data with both spatial and temporal structure. $\boldsymbol{U} \in \mathbb{R}^{625\times 2}$ are the spatial factors corresponding to a 25×25 imaging grid, with $\boldsymbol{u}_1 = \boldsymbol{U}_{\cdot 1}$ containing two non-overlapping regions of interest with smooth edges and

$\boldsymbol{u}_2 = \boldsymbol{U}_{\cdot 2}$ containing a single region of interest with sharp edges. $\boldsymbol{V} \in \mathbb{R}^{200\times 2}$ are the same temporal factors used in Section 4, namely time-localized sinusoidal pulses. These factors are shown in the top left panel of Figure 4.

Data are generated as $n = 200$ samples from the low-rank model $\boldsymbol{X} = \sum_{k=1}^{2} d_i \boldsymbol{u}_k \boldsymbol{v}_k^T + \boldsymbol{E}$ where the elements of $\boldsymbol{E}$ are are independently and identically drawn from a standard normal distribution. The signal-to-noise ratio is fixed at $d_1 = n/6$ and $d_2 = n/7$. As before, SFPCA is compared with several competing methods, including the two-way FPCA (TWFPCA) method of Huang *et al.* [8], the sparse SVD (SSVD) method of Lee *et al.* [26], the penalized matrix decomposition (PMD) of Witten *et al.* [5], and the sparse generalized PCA (SGPCA) of Allen *et al.* [3]. Each method was tuned according to the authors' recommendation, with SFPCA tuned using the greedy BIC scheme described above. For SFPCA and TWFPCA, $\boldsymbol{\Omega_u}$ is the second differences matrix over a 25×25 grid and $\boldsymbol{\Omega_v}$ is the second-differences matrix of a chain graph of length 200 (*i.e.*, a tridiagonal matrix with $(-1, 2, -1)$ on the tridiagonal). For SGPCA, the generalizing operators ($\boldsymbol{Q}, \boldsymbol{R}$ matrices) were again constructed from $\boldsymbol{\Omega_u}$ and $\boldsymbol{\Omega_v}$ using the methods suggested by Allen and Maletić-Savatić [27].

Qualitative results from this study are shown in Figure 4, where we see that SFPCA clearly outperforms the competing methods. Results for the temporal ($\boldsymbol{V}$) factors are similar to those for our one-way simulation, so we focus on the spatial ($\boldsymbol{U}$) factors here. The standard SVD provides neither sparsity, nor spatial smoothness, though the outline of the true signals can be discerned. TWFPCA recovers the smooth structure spatial signals well, but is not able to provide sparsity elsewhere. PMD appears to identify the signal, but as it does not allow for spatial smoothness, is insufficiently sparse elsewhere. SFPCA and optimally tuned SGPCA (here shown with $\sigma = 1$) both perform well here, but SGPCA is unable to recover the temporal smoothness patterns in the right singular vectors.

Quantitative results are presented in Table 2, where again we report the true positive rate (TP) and false positive rate (FP) for support recovery, as well as the relative angle and relative squared error to measure smoothness, which measure overall signal recovery. (See the main text for definitions). Consistent with the qualitative results, TWFPCA does well at recovering the true spatial signal in the first left singular vector, but cannot identify the sparse activation regions. Optimally-tuned SGPCA and SFPCA both perform well, with SFPCA slightly outperforming for the leading singular vectors and SGPCA outperforming for the following singular vectors. The good performance of SGPCA on this example is somewhat surprising as GPCA assumes smoothness in the noise, which is here IID, rather than the signal itself.

		TWFPCA	SSVD	PMD	SGPCA ($\sigma = 1$)	SGPCA ($\sigma = 5$)	SFPCA
$\boldsymbol{u}_1$	TP	-	**0.944** (.004)	0.697 (.005)	0.843 (.005)	0.532 (.004)	0.876 (.013)
	FP	-	0.611 (.111)	0.015 (.002)	0.024 (.002)	**0.000** (.000)	0.007 (.012)
	r$\angle$	**0.0832** (.041)	0.608 (.110)	0.934 (.024)	0.321 (.011)	1.140 (.034)	0.356 (.084)
$\boldsymbol{v}_1$	TP	-	**0.852** (.004)	0.679 (.004)	0.629 (.005)	0.659 (.007)	0.765 (.006)
	FP	-	0.617 (.111)	0.259 (.003)	**0.018** (.001)	0.045 (.004)	0.055 (.033)
	r$\angle$	0.252 (.071)	0.664 (.115)	0.565 (.009)	0.235 (.004)	0.186 (.005)	**0.142** (.053)
$\boldsymbol{u}_2$	TP	-	**0.892** (.005)	0.751 (.006)	0.679 (.006)	0.031 (.002)	0.562 (.016)
	FP	-	0.616 (.111)	0.202 (.005)	0.032 (.002)	**0.000** (.000)	0.006 (.011)
	r$\angle$	0.498 (.100)	0.547 (.105)	0.376 (.011)	**0.325** (.010)	3.650 (.088)	0.568 (.107)
$\boldsymbol{v}_2$	TP	-	**0.996** (.001)	0.983 (.003)	0.981 (.003)	0.659 (.010)	0.946 (.008)
	FP	-	0.614 (.111)	0.256 (.002)	**0.014** (.001)	0.058 (.005)	0.024 (.022)
	r$\angle$	0.720 (.120)	0.647 (.114)	0.439 (.007)	**0.213** (.006)	1.240 (.036)	0.355 (.084)
	rSE	0.276 (.001)	0.501 (.001)	0.470 (.004)	**0.203** (.003)	0.642 (.015)	0.212 (.001)

Table 2. Performance of various regularized PCA methods for the simulation study described in Section B.1. Results are averaged over 50 replicates, with standard errors given in parentheses. For each method, the true positive rate (TP), false positive rate (FP), relative angle compared to that of the SVD (r$\angle$), and relative squared error compared to that of the SVD (rSE) are reported. (TP and FP are not reported for the non-sparse TWFPCA.) The best performing method on each metric is bold-faced. Both SFPCA and SGPCA perform well on this example, though SFPCA has the additional advantage of not requiring the user to choose the smoothing parameter σ.

B.2. Additional EEG Results

In Section 5 of the main text, we compared the estimated SFPCA factors with ICA on electroencephalography (EEG) data from the UCI Machine Learning repository. In Figure 5, we show the results of applying (standard) PCA, two-way FPCA [8], two-way SPCA via the penalized matrix decomposition (TWSPCA) [5], and two-way sparse generalized PCA (TWSGPCA) [3].

As noted in the main body of the text, SFPCA and ICA identify similar temporal and temporal patterns for the first two components, but the SFPCA components have superior temporal sparsity, yielding improved interpretability. Standard PCA returns similar results to ICA, again failing to identify structure after the first two components. TWFPCA identifies smooth and biologically plausible smooth signals in all components, but cannot yield sparse estimates, hindering interpretation. TWSPCA returns similar first components (recall that these estimates are only defined up to a sign factor), but returns significantly more jagged estimates for the following components. The temporal components estimated by TWSGPCA are significantly more jagged and less sparse than those returned by SFPCA and do not exhibit meaningful temporal or spatial localization.

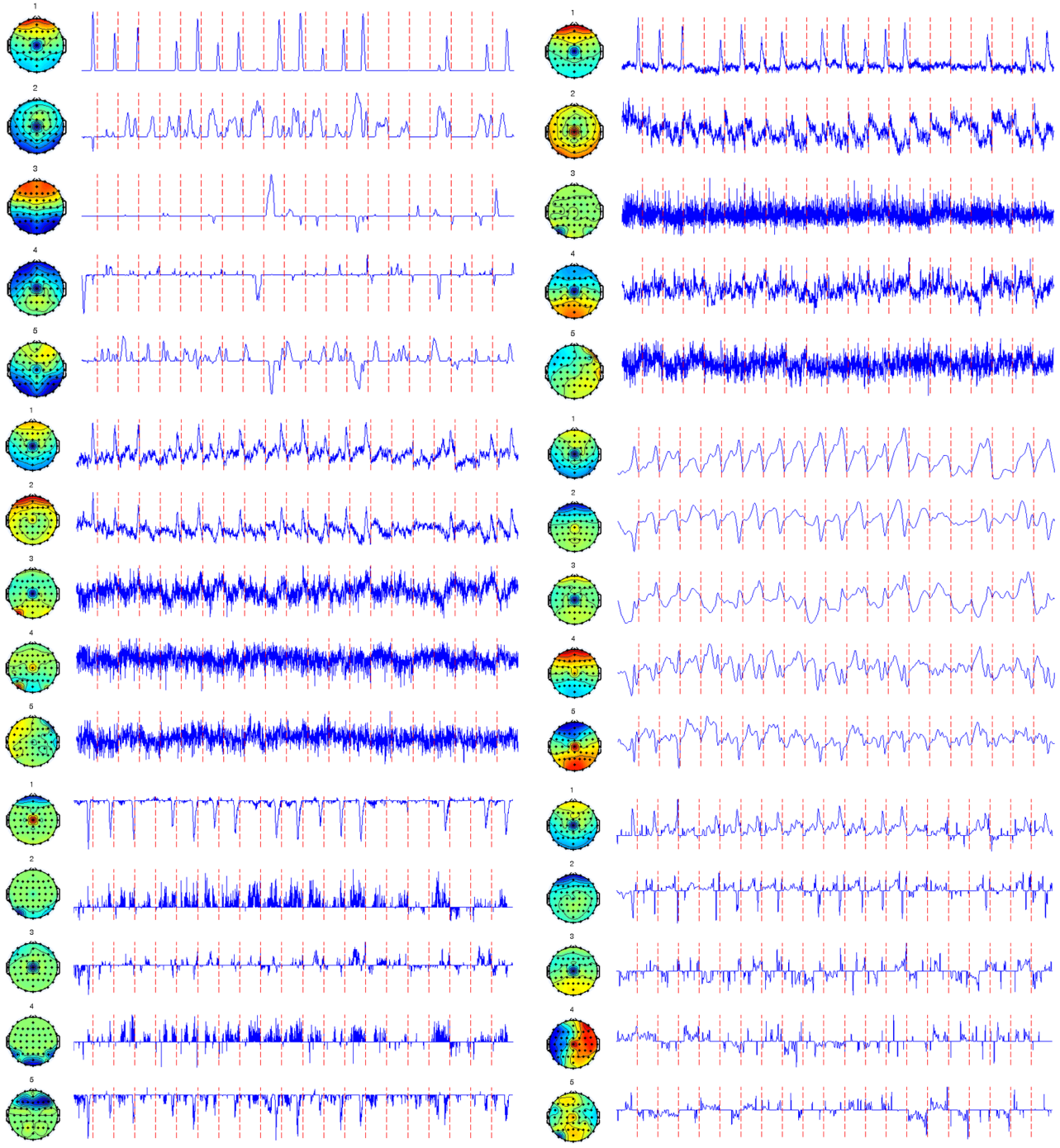

Fig. 5. EEG Case Study: first five spatial and temporal components as calculated by SFPCA (top-left), ICA (top-right), PCA (middle-left), TWFPCA (middle-right), TWSPCA / PMD (bottom-left), and TWSGPCA (bottom-right). The spatial and temporal signals identified by SFPCA are the most interpretable (sparsity) and the most biologically plausible (smoothness), while, unlike the other methods considered, also identifying meaningful structure past the first two components.

C. ADDITIONAL BACKGROUND

Since its introduction by Pearson [28] and Hotelling [29], principal component analysis (PCA) has been a mainstay of applied statistics. PCA provides a unified, computationally efficient, and mathematically elegant approach to dimension reduction, data visualization, and feature engineering. The usefulness of PCA has lead to its rediscovery by many other fields where it is variously known as the Karhunen-Loève transform [30], [31] in the theory of stochastic process, the method of empirical orthogonal functions in the environmental and atmospheric sciences (at least when the observation grid is regular; see the note of Buell [32] and the discussion thereof by Wikle and Cressie [33]), and the proper orthogonal decomposition in various engineering fields [34], among other names. In its classical form, PCA is performed on a (centered) data matrix $\boldsymbol{X}$ by taking the eigendecomposition of the scatter (covariance) matrix: $\boldsymbol{X}^T\boldsymbol{X}$. The Eckart-Young theorem [35], [36] establishes an equivalence between this formulation and the low-rank formulation we consider:

$$\boldsymbol{X} = \sum_{i=1}^{k} d_i \boldsymbol{u}_i \boldsymbol{v}_i \text{ for } \{\boldsymbol{u}_i\}_{i=1}^k, \{\boldsymbol{v}_i\}_{i=1}^k \text{ orthogonal elements of } \mathbb{R}^n, \mathbb{R}^p \text{ respectively.}$$

The elements of this low-rank representation can be found using the singular value decomposition of $\boldsymbol{X}$, which can be efficiently computed for large data matrices using the algorithms of Golub and Kahan [37] and Golub and Reinsch [38], as well as many more modern variants. We favor the low-rank formulation as it captures patterns in both the rows and columns of $\boldsymbol{X}$, but emphasize that equivalence between the two formulations holds only for the standard formulations and is broken when regularization is introduced. While we focus on matrix decomposition approaches, PCA may also be interpreted as the MLE of a certain probabilistic model, as shown by Tipping and Bishop [39]. The model-based framing is particularly useful when extending PCA to more complex data structures, *e.g.*, the integrative PCA (iPCA) model for multi-block data recently proposed by Tang and Allen [40].

When applying PCA, the selection of the true number of principal components is an important task. The "scree" heuristic of Cattell [41] considers the rate at which the singular values of $\boldsymbol{X}$ level off. To address the inherent subjectivity of this approach, several data-driven cross-validation-type techniques have been proposed [42]–[47]. More recently, rank selection methods based on the sampling distribution of noise eigenvalues have been proposed [48]–[50]. Several of these strategies are based on recent developments in random matrix theory which characterize the asymptotic properties of random matrices for standard null hypotheses [51], [52]. Before proceeding further, we note that we have only touched on a small fraction of the vast literature on PCA and refer the reader to the book of Jolliffe [53] or the more recent review of Abdi and Williams [54] for more a comprehensive coverage. The statistical properties of PCA have been studied by many authors, among which Anderson [55] and James [56] stand out as early and important references. These authors, and the rich theory developed afterward [57], establish asymptotic consistency of PCA in the large-sample ($n \to \infty$) setting. As statistical interest in data-sets for which the "aspect ratio" n/p is small grew, the short-comings of standard PCA were widely noted. New results in random matrix confirmed this observation: that standard PCA performs quite poorly unless the aspect ratio is large [58]–[62]. (Johnstone and Paul [63] give a useful and accessible review of the implications of random matrix theory for PCA. Bai and Ng [64] review the closely related literature on high-dimensional econometric factor models, among which Bai [65] stands out as a key reference.) To address this, regularized variants of PCA were proposed, several of which were later shown to yield consistent estimates.

The earliest forms of regularized PCA to appear in the literature arose in the functional data analysis community, where the principal components themselves were assumed to follow a smooth (functional) structure with respect to some norm. The early development of PCA of functional dates back to Dauxois *et al.* [66] and Besse and Ramsay [67], but Rice and Silverman [68] was the first, to the best of our knowledge, to explicitly impose a curvature penalty and propose an explicitly *functional* PCA (FPCA). Silverman [6] later penalized the curvature by altering the constraint region of the PCA problem and showed that this approach is equivalent to changing the norm and closely related to half-smoothing the data. From here, Huang *et al.* [7] showed that this approach can be formulated as a regression problem with a penalty on the $\boldsymbol{v}$-terms. Huang *et al.* [8] extended this idea to the low-rank model and proposed two-way FPCA via an alternating penalized regression scheme. They further established that this approach could be interpreted as attaining a (potentially local) solution to a penalized SVD problem. Zhang *et al.* [69] proposed a robust extension of the method of Huang *et al.* [8] where the Frobenius loss used for formulate the low-rank model is replaced by a robust loss function. Allen [20] later extended these approaches to the multi-way (tensor) setting. The literature on FPCA is vast and we refer the reader to the books by Ramsay and Silverman [70], [71] and the review of Hall [72] for more comprehensive coverage.

Sparsity-inducing regularized PCA (SPCA) was first proposed by Jolliffe *et al.* [73] who augmented the eigenvalue formulation of PCA with an ℓ_1 (LASSO [21]) constraint on the eigenvectors. Yuan and Zhang [74] and Ma [75] proposed algorithmic variants of the sparse eigenvalue problem which incorporate truncation and hard-thresholding steps, respectively, into standard eigenvector algorithms. (Journée *et al.* [76] give an interesting variant of this approach which retains convexity.) Convex semi-definite relaxations of the sparse eigenvalue problem were proposed by several authors [77]–[79], while Moghaddam *et al.* [80] propose a greedy search scheme. Johnstone and Lu [60] proposed a wavelet thresholding method which attempts to improve estimation of the covariance eigenstructure before standard PCA is performed, while Deshpande and Montanari [81] consider an algorithm based on direct covariance thresholding previously considered by Bickel and Levina [82], [83]. Wang *et al.* [84] propose an iterative approach to approximately solve the k-sparse eigenvalue problem with statistical guarantees. Finally, Asteris *et al.* [85] propose an intriguing method for estimating several sparse principal components based on bipartite graph matching.

We note that, however, that many of these approaches are derived from non-convex problems, which limits their theoretical tractability and computational efficiency. Additionally, these methods require instantiating and repeated use of the sample covariance matrix, which may be expensive for large scale problems. To address this, Zou *et al.* [86] proposed an alternative formulation which builds upon the ELASTICNET [22] penalized regression approach but requires solving a bi-convex problem using an iterative alternating regression scheme, an computational strategy shared with the low-rank model. Gataric *et al.* [87] propose an approach based on aggregating principal components of random low-dimensional projections of $\boldsymbol{X}$ which helps to limit the computational complexity. The majority of the methods discussed above identify only the leading principal component: if additional principal components are desired, they can be applied recursively to a "deflated" matrix.

The most commonly used deflation method is that of Hotelling ($\boldsymbol{X} := \boldsymbol{X} - d\boldsymbol{u}\boldsymbol{v}^T$) though Mackey [88] presents alternative approaches with better orthogonality properties. More recently, manifold optimization techniques have been used by Benidis *et al.* [89] and by Chen *et al.* [90] to simultaneously estimate multiple sparse principal components.

An early form of low-rank approximation with sparse factors was considered by Zhang *et al.* [91], [92]. In the statistical literature, the low-rank model used in our SFPCA formulation was first considered by Shen and Huang [4] for one-way sparsity and by Witten *et al.* [5] for two-way sparsity: both proposed alternating regression schemes to calculate leading singular values, though the Lagrangian form of Allen *et al.* [3] is closer to our approach. Lee *et al.* [26] and Yang *et al.* [93] proposed similar sparse singular value frameworks. Allen [94] extended the sparse low-rank model to the multi-way (tensor) setting. Udell *et al.* [95] review a range of similar models with the squared error loss replaced by other (exponential family) losses.

Many theoretical results for SPCA have been established in the literature, primarily for the covariance model: see, *e.g.*, the papers by Amini and Wainwright [96], Jung and Marron [97], [98], Vu and Lei [99], [100], Birnbaum *et al.* [101], Berthet and Rigollet [102], Shen *et al.* [103], d'Aspremont *et al.* [104], Cai *et al.* [105], Lei and Vu [106], Krauthgamer *et al.* [107], Ma and Wigderson [108], Wang *et al.* [109], and Bresler *et al.* [110], among many others. For that reason, we do not attempt to pain a comprehensive picture here, instead referring the reader to the review paper of Zou and Xue [111]. In addition to standard sparse PCA, several other sparse PCA variants have been proposed, including non-negative sparse PCA [27], [112], structured-sparse PCA [113], [114], sparse PCA with structured noise [3], contamination-robust sparse PCA [115], [116], and distributionally-robust sparse PCA [117], [118]. Lu *et al.* [119] propose an extension of sparse PCA to data sampled from an exponential family, building on an early proposal of Collins *et al.* [120]. (See also the related proposals of Lee *et al.* [121] and of Liu *et al.* [122].) Many of these schemes are based on the LASSO [21] penalty or structured variants thereof [113], [123], but the use of non-convex penalties has been occasionally considered: Shen and Huang [4] compare the use of SCAD [124] and hard-thresholding ("ℓ_0") penalties in their scheme, while Lee *et al.* [125] augment the method of Zou *et al.* [86] with a screening step followed by a penalized regression method using an ADAPTIVELASSO [126], SCAD [124], or MCP penalty [127].

The combination of sparsity and smoothness that we consider has not been extensively explored in the matrix factorization literature, though Slawski *et al.* [128] and Hebiri and Geer [129] explore similar ideas in a regression context. Based on an early draft of this paper, Li *et al.* [130] propose a framework for *supervised* SFPCA which combines standard (unsupervised) SFPCA with the Supervised SVD [131] and Mohammadi-Nejad *et al.* [132] propose a sparse and functional version of Canonical Correlation Analysis (CCA) [133]. Chen and Lei [134] propose a *localized* FPCA which performs FPCA with the additional constraint that the estimated factors have localized support, inducing similar sparsity structures to what we observe from SFPCA. While similar in name, the multilevel sparse functional PCA of Di *et al.* [135] refers to sparsely-sampled functional data, not sparsity in the factor loadings as we consider here.

D. ADDITIONAL REFERENCES

[1] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*. Athena Scientific, 2003, ISBN: 978-1-886-52945-8.

[2] M. Slater, "Lagrange multipliers revisited: A contribution to non-linear programming," Tech. Rep. Cowles Commission Discussion Paper No. 403 (Mathematics), 1950, Reissued as Cowles Foundation Discussion No. 80. Available at: http://cowles.yale.edu/sites/default/files/files/pub/d00/d0080.pdf.

[3] G. I. Allen, L. Grosenick, and J. Taylor, "A generalized least-square matrix decomposition," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 145–159, 2014. DOI: 10.1080/01621459.2013.852978.

[4] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1015–1034, 2008. DOI: 10.1016/j.jmva.2007.06.007.

[5] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009. DOI: 10.1093/biostatistics/kxp008.

[6] B. W. Silverman, "Smoothed functional principal components analysis by choice of norm," *Annals of Statistics*, vol. 24, no. 1, pp. 1–24, 1996. DOI: 10.1214/aos/1033066196.

[7] J. Z. Huang, H. Shen, and A. Buja, "Functional principal components analysis via penalized rank one approximation," *Electronic Journal of Statistics*, vol. 2, pp. 678–695, 2008. DOI: 10.1214/08-EJS218.

[8] ——, "The analysis of two-way functional data using two-way regularized singular value decompositions," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1609–1620, 2009. DOI: 10.1198/jasa.2009.tm08024.

[9] R. A. Poliquin and R. T. Rockafellar, "Generalized Hessian properties of regularized nonsmooth functions," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1121–1137, 1996. DOI: 10.1137/S1052623494279316.

[10] ——, "Prox-regular functions in variational analysis," *Transactions of the American Mathematical Society*, vol. 348, no. 5, pp. 1805–1838, 1996. DOI: 10.1090/S0002-9947-96-01544-9.

[11] P. Gong, C. Zhang, Z. Lu, J. Hang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *ICML 2013: Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds., vol. 28, Atlanta, Georgia, USA: PMLR, 2013, pp. 37–45. Available at: http://proceedings.mlr.press/v28/gong13a.html.

[12] Y.-L. Yu, "On decomposing the proximal map," in *NIPS 2013: Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26, Lake Tahoe, NV, USA, 2013, pp. 91–99. Available at: https://papers.nips.cc/paper/4863-on-decomposing-the-proximal-map.

[13] A. Beck, *First-order methods in optimization*, ser. MOS-SIAM Series on Optimization. 2017. DOI: 10.1137/1.9781611974997.

[14] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Mathematical Methods of Operations Research*, vol. 66, no. 3, pp. 373–407, 2007. DOI: 10.1007/s00186-007-0161-1.

[15] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1758–1789, 2013. DOI: 10.1137/120887795.

[16] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001. DOI: 10.1023/A:1017501703105.

[17] Y. Xu and W. Yin, "A globally convergent algorithm for nonconvex optimization based on block coordinate update," *Journal of Scientific Computing*, vol. 72, no. 2, pp. 700–734, 2017. DOI: 10.1007/s10915-017-0376-0.

[18] R. Mazumder, J. H. Friedman, and T. Hastie, "SparseNet: Coordinate descent with nonconvex penalties," *Journal of the American Statistical Association*, vol. 106, no. 495, pp. 1125–1138, 2011. DOI: 10.1198/jasa.2011.tm09738.

[19] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics*, vol. 5, no. 1, pp. 232–253, 2011. DOI: 10.1214/10-AOAS388.

[20] G. I. Allen, "Multi-way functional principal components analysis," in *CAMSAP 2013: Proceedings of the 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, V. Cevher and S. Gazor, Eds., St. Martin, France: IEEE, 2013, pp. 220–223. DOI: 10.1109/CAMSAP.2013.6714047.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[22] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005. DOI: 10.1111/j.1467-9868.2005.00503.x.

[23] R. J. Tibshirani and J. Taylor, "Degrees of freedom in lasso problems," *Annals of Statistics*, vol. 40, no. 2, pp. 1198–1232, 2012. DOI: 10.1214/12-AOS1003.

[24] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978. DOI: 10.1214/aos/1176344136.

[25] G. Claeskens and N. L. Hjort, *Model selection and model averaging*, 1st, ser. Cambridge Series in Statistical and Probabilistic Mathematics 27. Cambridge University Press, 2008, ISBN: 978-0-521-85225-8. DOI: 10.1017/CBO9780511790485.

[26] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, "Biclustering via sparse singular value decomposition," *Biometrics*, vol. 66, no. 4, pp. 1087–1095, 2010. DOI: 10.1111/j.1541-0420.2010.01392.x.

[27] G. I. Allen and M. Maletić-Savatić, "Sparse non-negative generalized PCA with applications to metabolomics," *Bioinformatics*, vol. 27, no. 21, pp. 3029–3035, 2011. DOI: 10.1093/bioinformatics/btr522.

[28] K. Pearson F.R.S., "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901. DOI: 10.1080/14786440109462720.

[29] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933. DOI: 10.1037/h0071325.

[30] K. Karhunen, "Zur spektraltheorie stochastischer prozesse," *Annales Academiae Scientiarum Fennicae*, vol. 34, pp. 1–7, 1946.

[31] M. Loève, "Fonctions aléatoires a décomposition orthogonale exponentielle," *La Revue Scientifique*, vol. 84, no. 3, pp. 159–161, 1946.

[32] C. E. Buell, "Integral equation representation for factor analysis," *Journal of the Atmospheric Sciences*, vol. 28, no. 1, pp. 1502–1505, 1971. DOI: 10.1175/1520-0469(1971)028<1502:IERFFA>2.0.CO;2.

[33] C. K. Wikle and N. Cressie, "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, vol. 86, no. 4, pp. 815–829, 1999. DOI: biomet/86.4.815.

[34] G. Berkooz, P. Holmes, and J. L. Lumley, "The proper orthogonal decomposition in the analysis of turbulent flows," *Annual Review of Fluid Mechanics*, vol. 25, pp. 539–575, 1993. DOI: 10.1146/annurev.fl.25.010193.002543.

[35] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936. DOI: 10.1007/BF02288367.

[36] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Review*, vol. 35, no. 4, pp. 551–566, 1993. DOI: 10.1137/1035134.

[37] G. H. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, vol. 2, no. 2, pp. 205–224, DOI: 10.1137/0702016.

[38] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, no. 5, pp. 403–420, DOI: 10.1007/BF02163027.

[39] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 61, no. 3, pp. 611–622, 1999. DOI: 10.1111/1467-9868.00196.

[40] T. M. Tang and G. I. Allen, "Integrative principal components analysis," *ArXiv Pre-Print 1810.00832*, 2018. Available at: http://arxiv.org/abs/1810.00832.

[41] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966. DOI: 10.1207/s15327906mbr0102_10.

[42] S. Wold, "Cross-validatory estimation of the number of components in factor and principal components models," *Technometrics*, vol. 20, no. 4, pp. 397–405, 1978. DOI: 10.1080/00401706.1978.10489693.

[43] H. T. Eastment and W. J. Krzanowski, "Cross-validatory choice of the number of components from a principal component analysis," *Technometrics*, vol. 24, no. 1, pp. 73–77, 1982. DOI: 10.1080/00401706.1982.10487712.

[44] A. Buja and N. Eyuboglu, "Remarks on parallel analysis," *Multivariate Behavioral Research*, vol. 27, no. 4, pp. 509–540, 1992. DOI: 10.1207/s15327906mbr2704_2.

[45] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001. DOI: 10.1093/bioinformatics/17.6.520.

[46] A. B. Owen and P. O. Perry, "Bi-cross-validation of the SVD and the nonnegative matrix factorization," *Annals of Applied Statistics*, vol. 3, no. 2, pp. 564–594, 2009. DOI: 10.1214/08-AOAS227.

[47] J. Josse and F. Husson, "Selecting the number of components in principal component analysis using cross-validation approximations," *Computational Statistics & Data Analysis*, vol. 56, no. 6, pp. 1869–1879, 2012. DOI: 10.1016/j.csda.2011.11.012.

[48] R. R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2625–2638, 2008. DOI: 10.1109/TSP.2008.917356.

[49] S. Kritchman and B. Nadler, "Determining the number of components in a factor model from limited noisy data," *Chemometrics and Intelligent Laboratory Systems*, vol. 94, no. 1, pp. 19–32, 2008. DOI: 10.1016/j.chemolab.2008.06.002.

[50] Y. Choi, J. Taylor, and R. Tibshirani, "Selecting the number of principal components: Estimation of the true rank of a noisy matrix," *Annals of Statistics*, vol. 45, no. 6, pp. 2590–2617, 2017. DOI: 10.1214/16-AOS1536.

[51] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001. DOI: 10.1214/aos/1009210544.

[52] D. Paul and A. Aue, "Random matrix theory in statistics: A review," *Journal of Statistical Planning and Inference*, vol. 150, pp. 1–29, 2014. DOI: 10.1016/j.jspi.2013.09.005.

[53] I. T. Jolliffe, *Principal component analysis*, 2nd, ser. Springer Series in Statistics. Springer-Verlag New York, 2002, ISBN: 978-0-387-95442-4. DOI: 10.1007/b98835.

[54] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews (WIRES): Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. DOI: 10.1002/wics.101.

[55] T. W. Anderson, "Asymptotic theory for principal component analysis," *Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 122–148, 1963. DOI: 10.1214/aoms/1177704248.

[56] A. T. James, "Distributions of matrix variates and latent roots derived from normal samples," *Annals of Mathematical Statistics*, vol. 35, no. 2, pp. 475–501, 1964. DOI: 10.1214/aoms/1177703550.

[57] R. J. Muirhead, *Aspects of multivariate statistical theory*, 2nd, ser. Wiley Series in Probability and Statistics. Wiley-Interscience, ISBN: 978-0-471-76985-9.

[58] J. Baik, G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *Annals of Probability*, vol. 33, no. 5, pp. 1643–1697, 2005. DOI: 10.1214/009117905000000233.

[59] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007. Available at: http://www3.stat.sinica.edu.tw/statistica/J17N4/J17N418/J17N418.html.

[60] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009. DOI: 10.1198/jasa.2009.0121.

[61] W. Wang and J. Fan, "Asymptotics of empirical eigenstructure for high dimensional spiked covariance," *Annals of Statistics*, vol. 45, no. 3, pp. 1342–1374, 2017. DOI: 10.1214/16-AOS1487.

[62] E. Dobriban, "Sharp detection in PCA under correlations: All eigenvalues matter," *Annals of Statistics*, vol. 45, no. 4, pp. 1810–1833, 2017. DOI: 10.1214/16-AOS1514.

[63] I. M. Johnstone and D. Paul, "PCA in high dimensions: An orientation," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1277–1292, 2018. DOI: 10.1109/JPROC.2018.2846730.

[64] J. Bai and S. Ng, "Large dimensional factor analysis," *Foundations and Trends® in Econometrics*, vol. 3, no. 2, pp. 89–163, 2008. DOI: 10.1561/0800000002.

[65] J. Bai, "Inferential theory for factor models of large dimensions," *Econometrica*, vol. 71, no. 1, pp. 135–171, 2003. DOI: 10.1111/1468-0262.00392.

[66] J. Dauxois, A. Pousse, and Y. Romain, "Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference," *Journal of Multivariate Analysis*, vol. 12, no. 1, pp. 136–154, 1982. DOI: 10.1016/0047-259X(82)90088-4.

[67] P. Besse and J. O. Ramsay, "Principal components analysis of sampled functions," *Psychometrika*, vol. 51, no. 2, pp. 285–311, 1986. DOI: 10.1007/BF02293986.

[68] J. A. Rice and B. W. Silverman, "Estimating the mean and covariance structure nonparametrically when the data are curves," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 53, no. 1, pp. 233–243, 1991. DOI: 10.1111/j.2517-6161.1991.tb01821.x.

[69] L. Zhang, H. Shen, and J. Z. Huang, "Robust regularized singular value decomposition with application to mortality data," *Annals of Applied Statistics*, vol. 7, no. 3, pp. 1540–1561, 2013. DOI: 10.1214/13-AOAS649.

[70] J. O. Ramsay and B. W. Silverman, *Applied functional data analysis: Methods and case studies*, 1st, ser. Spring Series in Statistics. Springer-Verlag New York, 2002, ISBN: 978-0-387-95414-1. DOI: 10.1007/b98886.

[71] ——, *Functional data analysis*, 2nd, ser. Springer Series in Statistics. Springer-Verlag New York, 2005, ISBN: 978-0-387-40080-8. DOI: 10.1007/b98888.

[72] P. Hall, "Principal component analysis for functional data: Methodology, theory, and discussion," in *The Oxford Handbook of Functional Data Analysis*, F. Ferraty and Y. Romain, Eds., 1st, Oxford University Press, 2011, pp. 210–235, ISBN: 978-0-199-56844-4. DOI: 10.1093/oxfordhb/9780199568444.013.8.

[73] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the LASSO," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 531–547, 2003. DOI: 10.1198/1061860032148.

[74] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *Journal of Machine Learning Research*, vol. 14, no. Apr, pp. 899–925, 2013. Available at: http://www.jmlr.org/papers/v14/yuan13a.html.

[75] Z. Ma, "Sparse principal component analysis and iterative thresholding," *Annals of Statistics*, vol. 41, no. 2, pp. 772–801, 2013. DOI: 10.1214/13-AOS1097.

[76] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 11, pp. 517–553, 2010. Available at: http://www.jmlr.org/papers/v11/journee10a.html.

[77] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Review*, vol. 49, no. 3, pp. 434–448, 2007. DOI: 10.1137/050645506.

[78] A. d'Aspremont, F. Bach, and L. El Ghaoui, "Optimal solutions for sparse principal component analysis," *Journal of Machine Learning Research*, vol. 9, pp. 1269–1294, 2008. Available at: http://www.jmlr.org/papers/v9/aspremont08a.html.

[79] V. Q. Vu, J. Cho, J. Lei, and K. Rohe, "Fantope projection and selection: A near-optimal convex relaxation of sparse PCA," in *NIPS 2013: Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., vol. 26, Lake Tahoe, NV, USA, 2013. Available at: https://papers.nips.cc/paper/5136-fantope-projection-and-selection-a-near-optimal-convex-relaxation-of-sparse-pca.

[80] B. Moghaddam, Y. Weiss, and S. Avidan, "Spectral bounds for sparse PCA: Exact and greedy algorithms," in *NIPS 2005: Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. C. Platt., Eds., 2005. Available at: https://papers.nips.cc/paper/2780-spectral-bounds-for-sparse-pca-exact-and-greedy-algorithms.

[81] Y. Deshpande and A. Montanari, "Sparse PCA via covariance thresholding," in *NIPS 2014: Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., vol. 27, Montréal, Canada, 2014. Available at: http://papers.nips.cc/paper/5406-sparse-pca-via-covariance-thresholding.

[82] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Annals of Statistics*, vol. 36, no. 1, pp. 199–227, 2008. DOI: 10.1214/009053607000000758.

[83] ——, "Covariance regularization by thresholding," *Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008. DOI: 10.1214/08-AOS600.

[84] Z. Wang, H. Lu, and H. Liu, "Tighten after relax: Minimax-optimal sparse PCA in polynomial time," in *NIPS 2014: Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., vol. 27, Montréal, Canada, 2014. Available at: http://papers.nips.cc/paper/5252-tighten-after-relax-minimax-optimal-sparse-pca-in-polynomial-time.

[85] M. Asteris, D. Papailiopoulos, A. Kyrillidis, and A. G. Dimakis, "Sparse PCA via bipartite matchings," in *NIPS 2015: Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Montréal, Canada, 2015. Available at: http://papers.nips.cc/paper/5901-sparse-pca-via-bipartite-matchings.

[86] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–288, 2006. DOI: 10.1198/106186006X113430.

[87] M. Gataric, T. Wang, and R. J. Samworth, "Sparse principal component analysis via random projections," *ArXiv Pre-Print 1712.05630*, 2017. Available at: http://arxiv.org/abs/1712.05630.

[88] L. Mackey, "Deflation methods for sparse PCA," in *NIPS 2008: Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2008, pp. 1017–1024. Available at: https://papers.nips.cc/paper/3575-deflation-methods-for-sparse-pca.

[89] 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. DOI: 10.1109/TSP.2016.2605073.

[90] S. Chen, S. Ma, L. Xue, and H. Zou, "An alternating manifold proximal gradient method for sparse PCA and sparse CCA," *ArXiv Pre-Print 1903.11576*, 2019. Available at: http://arxiv.org/abs/1903.11576.

[91] Z. Zhang, H. Zha, and H. Simon, "Low-rank approximations with sparse factors I: Basic algorithms and error analysis," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 3, pp. 706–727, 2002. DOI: 10.1137/S0895479899359631.

[92] ——, "Low-rank approximations with sparse factors II: Penalized methods with discrete Newton-like iterations," *SIAM Journal on Matrix Analysis and Applications*, vol. 25, no. 4, pp. 901–920, 2004. DOI: 10.1137/S0895479801394477.

[93] D. Yang, Z. Ma, and A. Buja, "A sparse singular value decomposition method for high-dimensional data," *Journal of Computational and Graphical Statistics*, vol. 23, no. 4, pp. 923–942, 2014. DOI: 10.1080/10618600.2013.858632.

[94] G. I. Allen, "Sparse higher-order principal components analysis," in *AISTATS 2012: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, vol. 22, Canary Islands, Spain: PMLR, 2012, pp. 27–36. Available at: http://proceedings.mlr.press/v22/allen12.html.

[95] M. Udell, C. Horn, R. Zadeh, and S. Boyd, "Generalized low rank models," *Foundations and Trends® in Machine Learning*, vol. 9, no. 1, 2016. DOI: 10.1561/2200000055.

[96] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," *Annals of Statistics*, vol. 37, no. 5B, pp. 2877–2921, 2009. DOI: 10.1214/08-AOS664.

[97] S. Jung and J. S. Marron, "PCA consistency in high dimension, low sample size context," *Annals of Statistics*, vol. 37, no. 6B, pp. 4104–4130, 2009. DOI: 10.1214/09-AOS709.

[98] T. T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," *Annals of Statistics*, vol. 41, no. 6, pp. 3074–3110, 2013. DOI: 10.1214/13-AOS1178.

[99] V. Q. Vu and J. Lei, "Minimax rates of estimation for sparse PCA in high dimensions," in *AISTATS 2012: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, vol. 22, Canary Islands, Spain: PMLR, 2012, pp. 1278–1286. Available at: http://proceedings.mlr.press/v22/vu12.html.

[100] ——, "Minimax sparse principal subspace estimation in high dimensions," *Annals of Statistics*, vol. 41, no. 6, pp. 2905–2947, 2013. DOI: 10.1214/13-AOS1151.

[101] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul, "Minimax bounds for sparse PCA with noisy high-dimensional data," *Annals of Statistics*, vol. 41, no. 3, pp. 1055–1084, 2013. DOI: 10.1214/12-AOS1014.

[102] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *Annals of Statistics*, vol. 41, no. 4, pp. 1780–1815, 2013. DOI: 10.1214/13-AOS1127.

[103] D. Shen, H. Shen, and J. S. Marron, "Consistency of sparse PCA in high dimension, low sample size contexts," *Journal of Multivariate Analysis*, vol. 115, pp. 317–333, 2013. DOI: 10.1016/j.jmva.2012.10.007.

[104] A. d'Aspremont, F. Bach, and L. El Ghaoui, "Approximation bounds for sparse principal component analysis," *Mathematical Programming*, vol. 148, no. 1-2, pp. 89–110, Dec. 2014. DOI: 10.1007/s10107-014-0751-7.

[105] T. Cai, Z. Ma, and Y. Wu, "Optimal estimation and rank detection for sparse spiked covariance matrices," *Probability Theory and Related Fields*, vol. 161, no. 3-4, pp. 781–815, 2015. DOI: 10.1007/s00440-014-0562-z.

[106] J. Lei and V. Q. Vu, "Sparsistency and agnostic inference in sparse PCA," *Annals of Statistics*, vol. 43, no. 1, pp. 299–322, 2015. DOI: 10.1214/14-AOS1273.

[107] R. Krauthgamer, B. Nadler, and D. Vilenchik, "Do semidefinite relaxations solve sparse PCA up to the information limit?" *Annals of Statistics*, vol. 43, no. 3, pp. 1300–1322, 2015. DOI: 10.1214/15-AOS1310.

[108] T. Ma and A. Wigderson, "Sum-of-squares lower bounds for sparse PCA," in *NIPS 2015: Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Montréal, Canada, 2015. Available at: https://papers.nips.cc/paper/5724-sum-of-squares-lower-bounds-for-sparse-pca.

[109] T. Wang, Q. Berthet, and R. J. Samworth, "Statistical and computational trade-offs in estimation of sparse principal components," *Annals of Statistics*, vol. 44, no. 5, pp. 1896–1930, DOI: 10.1214/15-AOS1369.

[110] G. Bresler, S. M. Park, and M. Persu, "Sparse PCA from sparse linear regression," in *NEURIPS 2018: Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Montréal, Canada, 2018. Available at: https://papers.nips.cc/paper/8291-sparse-pca-from-sparse-linear-regression.

[111] H. Zou and L. Xue, "A selective overview of sparse principal component analysis," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1311–1320, 2018. DOI: 10.1109/JPROC.2018.2846588.

[112] R. Zass and A. Shashua, "Nonnegative sparse PCA," in *NIPS 2006: Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds., vol. 19, Vancouver, Canada, 2006. Available at: https://papers.nips.cc/paper/3104-nonnegative-sparse-pca.

[113] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *AISTATS 2010: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, Y. W. Teh and M. Titterington, Eds., vol. 9, Sardinia, Italy: PMLR, 2010, pp. 366–373. Available at: http://proceedings.mlr.press/v9/jenatton10a.html.

[114] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Statistical Science*, vol. 27, no. 4, pp. 450–468, 2012. DOI: 10.1214/12-STS394.

[115] C. Croux, P. Filzmoser, and H. Fritz, "Robust sparse principal component analysis," *Technometrics*, vol. 55, no. 2, pp. 202–214, 2013. DOI: 10.1080/00401706.2012.727746.

[116] M. Hubert, T. Reynkens, E. Schmitt, and T. Verdonck, "Sparse PCA for high-dimensional data with outliers," *Technometrics*, vol. 58, no. 4, pp. 424–434, 2016. DOI: 10.1080/00401706.2015.1093962.

[117] F. Han and H. Liu, "Scale-invariant sparse PCA on high-dimensional meta-elliptical data," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 275–287, 2014. DOI: 10.1080/01621459.2013.844699.

[118] ——, "ECA: High-dimensional elliptical component analysis in non-Gaussian distributions," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 252–268, 2018. DOI: 10.1080/01621459.2016.1246366.

[119] M. Lu, J. Z.Huang, and X. Qian, "Sparse exponential family principal component analysis," *Pattern Recognition*, vol. 60, pp. 681–691, 2016. DOI: 10.1016/j.patcog.2016.05.024.

[120] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal components analysis to the exponential family," in *NIPS 2001: Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14, Vancouver, BC, Canada, 2001, pp. 617–642. Available at: https://papers.nips.cc/paper/2078-a-generalization-of-principal-components-analysis-to-the-exponential-family.

[121] S. Lee, J. Z. Huang, and J. Hu, "Sparse logistic principal components analysis for binary data," *Annals of Applied Statistics*, vol. 4, no. 3, pp. 1579–1601, 2010. DOI: 10.1214/10-AOAS327.

[122] L. T. Liu, E. Dobriban, and A. Singer, "ePCA: High-dimensional exponential family PCA," *Annals of Applied Statistics*, vol. 12, no. 4, pp. 2121–2150, 2018. DOI: 10.1214/18-AOAS1146.

[123] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006. DOI: 10.1111/j.1467-9868.2005.00532.x.

[124] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001. DOI: 10.1198/016214501753382273.

[125] Y. K. Lee, E. R. Lee, and B. U. Park, "Principal component analysis in very high-dimensional spaces," *Statistica Sinica*, vol. 22, no. 3, pp. 933–956, 2012. DOI: 10.5705/ss.2010.149.

[126] H. Zou, "The adapative lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006. DOI: 10.1198/016214506000000735.

[127] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010. DOI: 10.1214/09-AOS729.

[128] M. Slawski, W. zu Castell, and G. Tutz, "Feature selection guided by structural information," *Annals of Applied Statistics*, vol. 4, no. 2, pp. 1056–1080, 2010. DOI: 10.1214/09-AOAS302.

[129] M. Hebiri and S. van de Geer, "The smooth-lasso and other $\ell_1 + \ell_2$-penalized methods," *Electronic Journal of Statistics*, vol. 5, pp. 1184–1226, 2011. DOI: 10.1214/11-EJS638.

[130] G. Li, H. Shen, and J. Z. Huang, "Supervised sparse and functional principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 25, no. 3, pp. 859–878, 2016. DOI: 10.1080/10618600.2015.1064434.

[131] G. Li, D. Yang, A. B. Nobel, and H. Shen, "Supervised singular value decomposition and its asymptotic properties," *Journal of Multivariate Analysis*, vol. 146, pp. 7–17, 2016. DOI: 10.1016/j.jmva.2015.02.016.

[132] A.-R. Mohammadi-Nejad, G.-A. Hossein-Zadeh, and H. Soltanian-Zadeh, "Structured and sparse canonical correlation analysis as a brain-wide multi-modal data fusion approach," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1438–1448, 2017. DOI: 10.1109/TMI.2017.2681966.

[133] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936. DOI: 10.2307/2333955.

[134] K. Chen and J. Lei, "Localized functional principal component analysis," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 1266–1275, 2015. DOI: 10.1080/01621459.2015.1016225.

[135] C. Di, C. M. Crainiceanu, and W. S. Jank, "Multilevel sparse functional principal component analysis," *Stat*, vol. 3, no. 1, pp. 126–143, 2014. DOI: 10.1002/sta4.50.