Sparse and Functional Principal Components Analysis

Genevera I. Allen

Departments of Statistics and Electrical & Computer Engineering, Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital,
6100 Main St. MS-138, Houston, TX 77005 gallen@rice.edu

September 12, 2013

Abstract

Regularized principal components analysis, especially Sparse PCA and Functional PCA, has become widely used for dimension reduction in high-dimensional settings. Many examples of massive data, however, may benefit from estimating both sparse AND functional factors. These include neuroimaging data where there are discrete brain regions of activation (sparsity) but these regions tend to be smooth spatially (functional). Here, we introduce an optimization framework that can encourage both sparsity and smoothness of the row and/or column PCA factors. This framework generalizes many of the existing approaches to Sparse PCA, Functional PCA and two-way Sparse PCA and Functional PCA, as these are all special cases of our method. In particular, 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 our method on simulated data and a neuroimaging example on EEG data. This work provides a unified framework for regularized PCA that can form the foundation for a cohesive approach to regularization in high-dimensional multivariate analysis.

1 Introduction

Regularized principal component analysis (PCA) has become a popular multivariate analysis tool for dimension reduction, pattern recognition, compressed sensing, and visualization of big-data. Regularization in PCA was first introduced in the form of Functional PCA (FPCA) for structured data, data in which variables are associated with a location, time point, or frequency for example, to encourage smoothness of the PC factors with respect to the data structure (Rice and Silverman, 1991; Silverman, 1996; Huang et al., 2008). Later, regularization to encourage sparsity in the PC factors, Sparse PCA (SPCA), was introduced to perform automatic feature selection in high-dimensional settings (Jolliffe et al., 2003; Zou et al., 2006; d'Aspremont et al., 2007, 2008; Shen and Huang, 2008; Johnstone and Lu, 2009; Journée et al., 2010; Ma, 2013; Cai et al., 2012; Vu and Lei, 2012). Other extensions of regularized PCA include extensions for structured noise (Allen et al., 2011), structured-sparse penalties (Witten et al., 2009; Jenatton et al., 2009; Bach et al., 2012), and combinations of regularization and non-negativity (Zass and Shashua, 2007; Allen and Maletić-Savatić, 2011). More recently, several have proposed two-way extensions of these approaches which regularize both the row and column PC factors (Huang et al., 2009; Witten et al., 2009; Lee et al., 2010; Allen et al., 2011; Yang et al., 2011). While some have proposed combinations of smoothness and sparsity in other contexts (Hebiri and van de Geer, 2011; Bühlmann et al., 2011), this has not been explored in the context of matrix factorizations.

With many types of big-data, however, regularizing the PC factors to encourage both sparsity and smoothness of the row and/or column factors may be of interest. Consider spatio-temporal neuroimaging data, for example, where one seeks to find discrete brain regions (sparsity) that are spatially contiguous (smoothness) and their temporal activation patterns (smoothness and possibly sparsity). Sparse and smooth PC factors could be of interest for a variety of other big-data settings in which the data is structured, including image data, proteomics data such as NMR spectroscopy and mass spectrometry, genomics data such as copy number variation and methylation, and other spatio-temporal data including that of weather and climate studies. In this paper, we seek to develop an optimization framework for Sparse and Functional PCA (SFPCA) that permits flexible combinations of smoothness and sparsity of the row and column PC factors. In doing so, our goal is to generalize and unify much of the existing literature on regularized PCA.

Beyond the applied motivation, sparsity and smoothness of PC factors is desirable for many reasons. First, PCA has been shown to be inconsistent in high-dimensional settings (Johnstone and Lu, 2009; Jung and Marron, 2009). Several have shown, however, that FPCA (Silverman, 1996) and SPCA (Johnstone and Lu, 2009; Amini and Wainwright, 2009; Shen et al., 2012; Vu and Lei, 2012; Cai et al., 2012; Ma, 2013) lead to consistent estimates of the PC factors. Thus, a generalization of these approaches should also prove to be consistent in high-dimensional regimes. Second, as Sparse PCA is widely used in compressed sensing, sparse and functional PCA could be particularly useful in these settings, especially for structured image data. Next, both FPCA and SPCA have been shown to improve signal recovery and provide automatic feature selection, respectively. Thus, we can expect sparse and functional PCA to improve interpretability and visualization of the estimated PC factors. Finally, current methods for SPCA and FPCA are distinct, and we ask whether there exists a framework that unifies these approaches. Thus, developing such a framework to permit both smoothness and sparsity addresses a fundamental question that could have implications for many other aspects of regularization in multivariate analysis.

In this paper, we make several major contributions: (1) We present a coherent optimization framework that permits flexible combinations of regularization to encourage smoothness and sparsity of the row and column PC factors. (2) We show that our proposed optimization framework unifies much of the existing literature on regularized PCA, where PCA, SPCA, FPCA, two-way SPCA, and two-way FPCA are all special cases of our approach. (3) We provide a tractable algorithm to fit our SFPCA method in big-data settings and provide strong algorithmic convergence guarantees. Because of these contributions, our SFPCA method enjoys many advantages over existing approaches to regularized PCA: (i) As our framework unifies many existing methods, our method inherits the desirable properties of both SPCA and FPCA, including improved signal recovery, automatic feature selection, and improved interpretability. (ii) Because we formulate a single optimization problem that permits both sparsity and smoothness, our flexible method is conducive to data-driven determination of the most appropriate types and amounts of regularization in the context of dimension reduction. (iii) By smoothing the factors, our method yields substantial improvements in feature selection over existing methods, as smoothness allows us to borrow strength across neighboring variables. Overall, this paper addresses a fundamental question in regularized PCA, and proposes a method that has significant advantages for dimension reduction of big-data.

This paper is organized as follows. We begin in Section 2.1 with background on regularization approaches to FPCA, SPCA and two-way extensions of these. In this paper, we specifically focus on the popular and well studied penalized regression approaches to regularized PCA (Zou et al., 2006; Huang et al., 2008; Shen and Huang, 2008; Witten et al., 2009; Huang et al., 2009; Lee et al., 2010; Allen et al., 2011) which use penalties such as the ℓ_1 or elliptical ℓ_2 norms to encourage sparsity or smoothness. Given these, one may presume that sparsity and smoothness can be achieved by simply adding additional ℓ_1 or ℓ_2 penalties to these existing frameworks. We explore this possibility in Section 2.2, and show that formulating an SFPCA model with desirable theoretical, numerical, and computational properties is not as simple as arbitrarily adding penalties to existing techniques. In Section 3, we introduce our optimization framework for SFPCA, proving that this generalizes many existing approaches to regularized PCA and has desirable numerical properties. We provide an efficient computational algorithm for SFPCA with a variety of possible sparse penalties in Section 4, providing strong convergence guarantees. Finally, we numerically test our method on simulated data in Section 5.1 and real EEG data in Section 6, concluding with a discussion in Section 7.

2 Background & Challenges with Regularized PCA

2.1 Background

Suppose we observe a data matrix, $\mathbf{X} \in \mathbb{R}^{n \times p}$, and assume that the data arises from a low-rank model: $\mathbf{X} = \sum_{k=1}^{K} d_k \, \mathbf{u}_k \, \mathbf{v}_k^T + \boldsymbol{\epsilon}$, where $\epsilon_{ij} \overset{iid}{\sim}$ with mean 0, $\mathbf{u}_k$ is the n length k^{th} left factor, and $\mathbf{v}_k$ is the p length k^{th} right factor. (Note that throughout this paper, we will assume that the data has previously been centered). Thus, we assume a common PCA model, but additionally assume that each row and column factor has sparse support and is smooth with respect to some known structure. For example, we assume that the rows and/or columns of $\mathbf{X}$ arise as discretized functional curves.

While there are many options for using regularization to encourage smoothness, we limit our consideration in this paper to the types of generalized ℓ_2 penalties widely used in smoothing splines, reproducing kernel Hilbert spaces, and functional data analysis as described in Ramsay (2006). In particular for continuous functional curves, penalizing the second or fourth derivatives, penalizes the curvature and hence encourages the estimated function to be smooth. For discretized functional curves, this can be generalized to penalizing the second or fourth differences between adjacent

variables, a penalty that can be written as a quadratic form: $\mathbf{x}^T \boldsymbol{\Omega} \mathbf{x}$, where $\boldsymbol{\Omega}$ is the positive semi-definite second or fourth differences matrix (Rice and Silverman, 1991; Ramsay, 2006).

Functional PCA was first proposed by Rice and Silverman (1991) by penalizing the curvature of the factors in the objective as described above; Silverman (1996) later penalized the curvature by altering the constraint region of the PCA problem and showed that this approach was equivalent to changing the norm and related to half-smoothing the data. Recently, Huang et al. (2008) showed that this approach can be formulated as a penalized regression problem. Then in Huang et al. (2009), this approach was extended to two-way FPCA by proposing alternating penalized regression problems that were related to two-way half-smoothing and could also be formulated as a coherent optimization problem in terms of quadratic penalties on $\boldsymbol{\Omega}_\mathbf{u}$ and $\boldsymbol{\Omega}_\mathbf{v}$. There are numerous proposals for Sparse PCA; Jolliffe et al. (2003); Zou et al. (2006) were among the first suggested approaches that employed ℓ_1 penalties to encourage sparsity. Later, d'Aspremont et al. (2007) proposed an approach using semi-definite programming, and Johnstone and Lu (2009) proposed to use thresholded wavelets. Recently, Shen and Huang (2008) suggested a popular penalized regression-based approach to SPCA. This approach has formed the basis of several suggested formulations of two-way SPCA via alternating penalized regressions (Witten et al., 2009; Lee et al., 2010; Allen et al., 2011). Given the popularity of and computational tractability of these approaches for SPCA, FPCA and two-way extensions of these, our development of sparse and functional PCA will be in the flavor of these methods for regularized PCA via alternating regressions. Additionally, these are the only optimization frameworks in the current literature that permit two-way regularization of the PC factors.

One may question, given all of these existing approaches for two-way FPCA or two-way SPCA, whether one cannot simply define a heuristic using these techniques to achieve the desired result of smooth and sparse PC factors. For example, why not smooth the data matrix first and then apply two-way SPCA? While this approach may provide some smoothness with sparse factors, it separates the smoothing from the low-rank estimation; thus, the resulting factors are not as smooth as in FPCA approaches, and it does not accommodate different amounts of smoothness on each of the factors, a significant disadvantage. Alternatively as illustrated in Huang et al. (2009), the two-way FPCA method is equivalent to half-smoothing the data, taking the SVD, and then half-smoothing the singular vectors. Why not simply apply two-way SPCA instead of the SVD in this process? While natural to do, half-smoothing the resulting SPCs would destroy the sparsity structure. On the other hand, one could take the two-way FPCA factors and then threshold them to achieve sparsity. The thresholded factors, however, no longer jointly determine a low-rank matrix factorization, as the left factor needs to be determined as a function of the right and vice versa. Finally, as any heuristic defined does not solve a coherent objective, the properties of the resulting solution are harder to study, and thus these are mathematically less appealing.

2.2 Challenges with Formulating Sparse and Functional PCA

Given the previous background and discussion, we seek an optimization framework to achieve both sparse and smooth row and column PC factors. As is most common in the regularized PCA literature, we will limit our consideration to estimating a single-rank model one at a time, $d\,\mathbf{u}\,\mathbf{v}^T$. Also, as we wish to regularize both the row and column PC factors, we consider frameworks with four penalty terms: $\lambda_\mathbf{u} P_\mathbf{u}(\mathbf{u})$ and $\lambda_\mathbf{v} P_\mathbf{v}(\mathbf{v})$ which are penalties that encourage sparsity or structured-sparsity, and $\alpha_\mathbf{u}\,\mathbf{u}^T\,\boldsymbol{\Omega}_\mathbf{u}\,\mathbf{u}$ and $\alpha_\mathbf{v}\,\mathbf{v}^T\,\boldsymbol{\Omega}_\mathbf{v}\,\mathbf{v}$ to encourage smoothness. Here, the λ's and α's are non-negative penalty parameters controlling the amount of sparsity and smoothness respectively. While there are many possible penalties to encourage sparsity, for this section, we will work with a simple and popular example: the ℓ_1-norm penalty, a direct convex relaxation of the ℓ_0-norm. Also as previously discussed, $\boldsymbol{\Omega}_\mathbf{u} \succeq 0$ is a $n \times n$ and $\boldsymbol{\Omega}_\mathbf{v} \succeq 0$ a $p \times p$ matrix that penalizes the roughness in $\mathbf{u}$ and $\mathbf{v}$, for example, the second or fourth differences matrix.

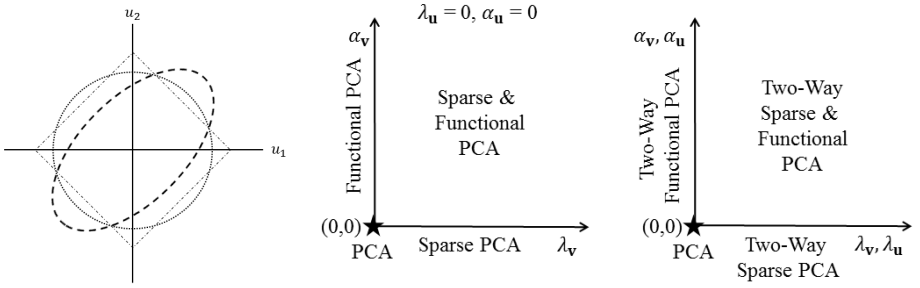
With this setup of ℓ_1 penalties to encourage sparsity and elliptical ℓ_2 penalties to encourage smoothness, we suggest that the ideal optimization model for SFPCA should satisfy the following three objectives:

1. Generalize much of the existing regularized PCA literature with PCA, SPCA, FPCA, and two-way SPCA and FPCA all special cases of the model when the corresponding regularization parameters are inactive.

2. Achieve desirable numerical properties including identifiability, non-degenerate solutions, well-scaled solutions, and balanced regularization.

3. Be conducive to tractable computational schemes for big-data settings.

Figure 1(b) illustrates the first objective which is necessary for formulating a method which permits data-driven determination of the types and amounts of regularization appropriate for unsupervised modeling of low-rank structure. Good numerical properties leading to identifiable, non-degenerate, and well-scaled solutions for the row and column



(a) ℓ_1, ℓ_2, and ellipsoidal ℓ_2 balls. (b) Illustration of Theorem 1.

Figure 1: Illustration of SFPCA regularization.

factors are of obvious importance. When working with four potential regularization terms, we suggest that balanced regularization, meaning that the solution must depend on all active (non-zero) penalty parameters and that increasing λ or α increases the amount of sparsity or smoothness respectively, is also desirable. This allows for ease of interpretation and data-driven selection of the penalty parameters as well as precludes regularization masking, an item discussed subsequently. Finally, as we are motivated by structured big-data such as in neuroimaging and climate studies, scalable computational algorithms are necessary for these applications.

Given the many existing approaches to two-way Sparse PCA and two-way Functional PCA, one may ask whether we can simply add elliptical ℓ_2 or ℓ_1 penalties respectively to these existing optimization frameworks to achieve a model for SFPCA. In the following discussion, we will show that three such logical extensions of existing approaches, while natural, have major flaws that do not satisfy our three objectives.

The rank-one SVD, and hence PCA, is the solution to minimizing the Frobenius norm loss between $\mathbf{X}$ and a rank-one matrix, $\mathbf{u}\,\mathbf{v}^T$. Hence, why not directly add ℓ_1 penalties and smoothing ℓ_2 penalties to this Frobenius norm loss?

$$\underset{\mathbf{u}(:\mathbf{u}^T\,\mathbf{u}\leq 1),\mathbf{v}(:\mathbf{v}^T\,\mathbf{v}\leq 1)}{\text{minimize}} \; || \mathbf{X} - (d)\,\mathbf{u}\,\mathbf{v}^T ||_F^T + \lambda_{\mathbf{u}} || \mathbf{u} ||_1 + \lambda_{\mathbf{v}} || \mathbf{v} ||_1 + \alpha_{\mathbf{u}}\,\mathbf{u}^T\,\mathbf{\Omega}_{\mathbf{u}}\,\mathbf{u} + \alpha_{\mathbf{v}}\,\mathbf{v}^T\,\mathbf{\Omega}_{\mathbf{v}}\,\mathbf{v}\,. \tag{1}$$

Also, consider the optimization problem for two-way FPCA proposed by Huang et al. (2009). As this gives two-way smoothness, why not add ℓ_1 penalties to additionally encourage sparsity?

$$\underset{\mathbf{u}(:\mathbf{u}^T\,\mathbf{u}\leq 1),\mathbf{v}(:\mathbf{v}^T\,\mathbf{v}\leq 1)}{\text{maximize}} \; \mathbf{u}^T\,\mathbf{X}\,\mathbf{v} - \frac{1}{2}\,\mathbf{u}^T(\mathbf{I} + \alpha_{\mathbf{u}}\,\mathbf{\Omega}_{\mathbf{u}})\,\mathbf{u}\,\mathbf{v}^T(\mathbf{I} + \alpha_{\mathbf{v}}\,\mathbf{\Omega}_{\mathbf{v}})\,\mathbf{v} - \lambda_{\mathbf{u}} || \mathbf{u} ||_1 - \lambda_{\mathbf{v}} || \mathbf{v} ||_1 \tag{2}$$

Finally, consider the optimization problem for two-way SPCA proposed by Allen et al. (2011) which is the Lagrange form of that proposed by Witten et al. (2009). As this gives two-way sparsity, why not add smoothing ℓ_2 penalties to additionally encourage smoothness?

$$\underset{\mathbf{u}(:\mathbf{u}^T\,\mathbf{u}\leq 1),\mathbf{v}(:\mathbf{v}^T\,\mathbf{v}\leq 1)}{\text{maximize}} \; \mathbf{u}^T\,\mathbf{X}\,\mathbf{v} - \lambda_{\mathbf{u}} || \mathbf{u} ||_1 - \lambda_{\mathbf{v}} || \mathbf{v} ||_1 - \alpha_{\mathbf{u}}\,\mathbf{u}^T\,\mathbf{\Omega}_{\mathbf{u}}\,\mathbf{u} - \alpha_{\mathbf{v}}\,\mathbf{v}^T\,\mathbf{\Omega}_{\mathbf{v}}\,\mathbf{v}\,. \tag{3}$$

We first examine these three natural extensions in terms of the second objective, desirable numerical properties. To begin with, consider (1), (2), and (3) without the scale constraints on the factors $\mathbf{u}$ and $\mathbf{v}$. The first terms of these objectives are invariant to scale transformations, meaning that $\hat{\mathbf{u}}c$ and $\hat{\mathbf{v}}/c$ give the same objective values for some $c > 0$. Thus, without scale constraints, $\mathbf{u}$ and $\mathbf{v}$ are not identifiable. Additionally, these problems are degenerate in that $\hat{\mathbf{u}} \to 0$ and $\hat{\mathbf{v}} \to \infty$ if the penalty parameters are such that $\mathbf{u}$ is regularized more than that of $\mathbf{v}$. A discussion of this phenomenon for two-way regularized PCA problems is given in Huang et al. (2009). This leads us to consider adding scale constraints. Typically PCA constrains the norm of the factors to be one. But as Witten et al. (2009) notes, norm constraints are non-convex and thus a relaxation, $\mathbf{u}^T\,\mathbf{u} \leq 1$ is considered.

Adopting these scale constraints, the first term of (1) needs a scale factor, $d > 0$ yielding $d\,\mathbf{u}\,\mathbf{v}^T$. Otherwise, the magnitude of the regularization parameters leading to sparse and/or smooth solutions will depend on the magnitude of the true scaling factor d, violating our principal of balanced regularization. With the scale-factor, d in (1), however, it is easy to see that the solution is still degenerate: Take $d\,\mathbf{u}\,\mathbf{v}^T$ to be the SVD solution. Then, we can take $\hat{d} = d/\epsilon^2$, $\hat{\mathbf{u}} = \mathbf{u}\,\epsilon$, and $\hat{\mathbf{v}} = \mathbf{v}\,\epsilon$ for some $\epsilon > 0$. When the left and right factors are regularized, the solution will then be numerically unstable with the scale term, d, going to infinity.

Now consider (2) and (3) with the convex scale constraints. These optimization problems essentially have double ℓ_2 regularization with an ℓ_2 ball from the scale-constraints and an ℓ_2 ellipsoidal ball corresponding to the smoothing regularizers. Double ℓ_2 regularization, however, leads to regularization masking, where for certain values of α, the smoothness regularization is rendered inactive as it is masked by the ℓ_2-norm scale constraint. This is depicted in Figure 1(a) where careful selection of λ, controlling the size of the ℓ_1-norm ball, and α, controlling the size of the elliptical ℓ_2 ball, is needed to find a sparse and smooth solution that intersects the ℓ_2-norm ball corresponding to the scale constraint.

Hence, we have shown that formulating an optimization model that achieves the desirable numerical properties outlined above is not as simple as adding regularizers to existing two-way SPCA or FPCA approaches. From this discussion, we have learned that controlling the scale of the factors is necessary as well as avoiding double ℓ_2 regularization and the associated regularization masking. Finally, while we have focused on the flaws of these logical extensions in terms of our second objective, careful examination of (1) reveals that SPCA and FPCA are not special cases and examination of (2) and (3) also reveal that these optimization problems are not suited to simple and computationally efficient optimization strategies such as the alternating regressions approach employed in numerous regularized PCA methods (Zou et al., 2006; Shen and Huang, 2008; Huang et al., 2008; Witten et al., 2009; Lee et al., 2010; Allen et al., 2011).

3 Sparse and Functional PCA Optimization Model

Our objective is to develop an optimization framework that permits flexible combinations of sparsity and/or smoothness of the PC row and column factors and that (1) generalizes much of the existing literature on regularized PCA, (2) is identifiable and has desirable numerical properties leading to non-degenerate, well-scaled, and balanced solutions, and (3) will lead to tractable computational approaches for big-data settings. We, then propose the following optimization problem that defines our SFPCA model:

$$\begin{aligned} \underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \quad & \mathbf{u}^T \mathbf{X}\, \mathbf{v} - \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) - \lambda_{\mathbf{v}} P_{\mathbf{v}}(\mathbf{v}) \\ \text{subject to} \quad & \mathbf{u}^T \left(\mathbf{I} + \alpha_{\mathbf{u}}\, \mathbf{\Omega}_{\mathbf{u}}\right) \mathbf{u} \leq 1 \;\&\; \mathbf{v}^T \left(\mathbf{I} + \alpha_{\mathbf{v}}\, \mathbf{\Omega}_{\mathbf{v}}\right) \mathbf{v} \leq 1. \end{aligned} \tag{4}$$

As in the previous section, $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}} \geq 0$ are regularization parameters controlling the amount of sparsity in $\mathbf{u}$ and $\mathbf{v}$ according to penalties, $P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ respectively; $\alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} \geq 0$ are regularization parameters controlling the smoothness of $\mathbf{u}$ and $\mathbf{v}$ according to the structure encoded in $\mathbf{\Omega}_{\mathbf{u}}$ and $\mathbf{\Omega}_{\mathbf{v}}$ respectively. We make some assumptions on our smoothness and sparsity penalties:

Assumption 1 (A1)**.** *$\mathbf{\Omega}_{\mathbf{u}} \succeq 0$ and $\mathbf{\Omega}_{\mathbf{v}} \succeq 0$.*

Assumption 2 (A2)**.** *$P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ are such that $P() : \mathbb{R} \to \mathbb{R}^+$ and are positive, homogeneous of order one, meaning that $P(cx) = cP(x)\ \forall\ c > 0$.*

Assumption 3 (A3)**.** *If $P_{\mathbf{u}}()$ or $P_{\mathbf{v}}()$ is non-convex, then $P()$ can be decomposed into the difference of two convex functions, $P(x) = P_1(x) - P_2(x)$ for $P_1()$ and $P_2()$ convex.*

These assumptions placed on our sparsity penalties are in fact quite general, encompassing nearly all the popular penalties in the literature; examples satisfying A2 and A3 include convex, sparse penalties such as the ℓ_1-norm, convex structured-sparse penalties such as the fused lasso (Tibshirani et al., 2005) or the group lasso (Yuan and Lin, 2006), and non-convex sparse penalties such as the SCAD penalty (Fan and Li, 2001) among many others as outlined in Gong et al. (2013). Notice that instead of adding penalties to enforce smoothness to the objective as in FPCA approaches (Huang et al., 2009), we incorporate these in the constraint region, which as we will see, permits balanced smoothness regularization, provides a generalization of existing regularized PCA methods, and leads to a simple and scalable computational strategy.

The following result shows that (4) satisfies our first objective, providing a general and unified framework for regularized PCA.

Theorem 1. *Assume A1-A3 and let $(\mathbf{u}^*, \mathbf{v}^*)$ be the optimal points of* (4)*. Then, the following hold:*

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

(ii) If $\lambda_{\mathbf{u}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then $\mathbf{u}^*$ and $\mathbf{v}^*$ are equivalent to the SPCA solution of Shen and Huang (2008).

(iii) If $\alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, then $\mathbf{u}^*$ and $\mathbf{v}^*$ are equivalent to a special case of the two-way SPCA solution in Allen et al. (2011), which is the Lagrangian of Witten et al. (2009) and closely related to that of Lee et al. (2010).

(iv) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}} = 0$, then $\mathbf{u}^*$ and $\mathbf{v}^*$ are equivalent to the FPCA solution of Silverman (1996) and Huang et al. (2008).

(v) If $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}} = 0$, then $\mathbf{u}^*$ and $\mathbf{v}^*$ are equivalent to the two-way FPCA solution of Huang et al. (2009).
(For parts (ii) and (iii), equivalencies hold for the appropriate $P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ employed in the referenced papers.)

Thus, for various combinations of the regularization parameters, our framework can return the solution to PCA (SVD), SPCA, FPCA, two-way SPCA, and two-way FPCA. When both the smoothness parameters, α, and the sparsity parameters λ, are non-zero, the solution to (4) will have both sparse and smooth components. This is depicted in Figure 1(b). Because of our general formulation, our method is conducive to data-driven determination of the amount of sparsity and/or smoothness that is appropriate for modeling both the left and right factors.

In the following set of properties and results, we show that our problem satisfies our second objective; namely, it is identifiable and satisfies the desirable numerical properties outlined in the previous section:

Property 1. *In the SFPCA problem* (4)*, $\mathbf{u}$ and $\mathbf{v}$ are identifiable up to a sign change.*

Proposition 1. *Assume A1-A3 and let $(\mathbf{u}^*, \mathbf{v}^*)$ be the optimal points of* (4)*.*

(i) There exists some value $\lambda_{\mathbf{u}}^{max}$ for which the solution to (4) *is $\mathbf{u}^* = \mathbf{0}$, and analogously some value $\lambda_{\mathbf{v}}^{max}$ at which $\mathbf{v}^* = \mathbf{0}$.*

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

(iii) Either $\mathbf{u}^{*,T}(\mathbf{I} + \alpha_{\mathbf{u}} \mathbf{\Omega}_{\mathbf{u}}) \mathbf{u}^* = 1$ and $\mathbf{v}^{*,T}(\mathbf{I} + \alpha_{\mathbf{v}} \mathbf{\Omega}_{\mathbf{v}}) \mathbf{v}^* = 1$ or $|| \mathbf{u}^* || = 0$ and $|| \mathbf{v}^* || = 0$, with the latter occurring when $\lambda_{\mathbf{u}} \geq \lambda_{\mathbf{u}}^{max}$ or $\lambda_{\mathbf{v}} \geq \lambda_{\mathbf{v}}^{max}$.

These numerical properties combine to show that our formulation of an SFPCA model avoids the identifiability, degeneracy, and un-balanced regularization of the naive regularization approaches previously discussed. Specifically, zero solutions can only occur if the sparsity parameters are sufficiently large, and otherwise the solutions are non-degenerate and well-scaled. Also, constraining the factors with an elliptical ℓ_2 ball avoids regularization masking; we are thus assured that the solution factors depend on all non-zero regularization parameters and increasing any one of these parameters increases the amount of smoothness or sparsity imposed.

Besides the generality and desirable numerical properties of our SFPCA model, there are further extensions of interest as outlined in the following remarks:

Remark 1. *Our optimization framework can easily be extended to the case of functional curves. Consider a two-way continuous random process $X(s,t)$ in domain $\mathcal{S} \times \mathcal{T}$, for example, a spatio-temporal process; then,* (4) *for continuous PC factors $u(s)$ and $v(t)$ can be written as:*

$$\underset{u(s),v(t)}{\text{maximize}} \int_{\mathcal{S}} \int_{\mathcal{T}} X(s,t)u(s)v(t)dsdt - \lambda_u \int_{\mathcal{S}} |u(s)|ds - \lambda_v \int_{\mathcal{T}} |v(t)|dt$$
$$\textit{subject to} \int_{\mathcal{S}} \left(u^2(s) + \alpha_u (u''(s))^2\right) ds = 1 \;\; \& \;\; \int_{\mathcal{T}} \left(v^2(t) + \alpha_v (v''(t))^2\right) dt = 1.$$

This optimization problem can be seen as an extension of the continuous FPCA methods discussed in Silverman (1996); Ramsay (2006).

Remark 2. *Our SFPCA optimization framework can be extended to encompass non-negative constraints:*

$$\text{maximize} \;\; (4) \;\; \textit{subject to} \;\; \mathbf{u} \geq 0, \;\& \;\; \mathbf{v} \geq 0, \tag{5}$$

as employed in non-negative PCA methods such as Zass and Shashua (2007); Allen and Maletić-Savatić (2011).

Finally, a motivation for regularizing factors in PCA is the fact that estimated PC directions are often inconsistent for the population version in high-dimensional settings (Johnstone and Lu, 2009; Jung and Marron, 2009). Two special cases of our model, the Sparse PCA approach of Shen and Huang (2008) and the Functional PCA approach of Silverman (1996) have been shown to consistent in these settings (Shen et al., 2012). Our unified optimization framework provides the foundation for studying general consistency properties of regularized PCA in high-dimensional settings, but is beyond the scope of this paper and left for future work.

4 Sparse and Functional PCA Algorithm

We seek to develop a tractable algorithm to optimize the SFPCA model, (4), in big-data settings. As noted in previous sections and as is common throughout the regularized PCA literature, we consider estimating a single-rank model at a time in a greedy manner, also called a deflation method and similar to the power method for computing eigenvalues (Witten et al., 2009; Journée et al., 2010). To estimate the rank-one model, (4) is estimated for $\mathbf{X}$; then $\mathbf{X}$ is deflated by the estimated rank-one solution, $\hat{\mathbf{X}}^{(1)} = \mathbf{X} - \hat{d}_1 \hat{\mathbf{u}}_1 \hat{\mathbf{v}}_1^T$ and (4) is fit to $\hat{\mathbf{X}}^{(1)}$ to estimate the rank-2 factors, and so forth. We use this so-called subtraction (Hotelling's) deflation in our numerical studies, but note that several other deflation schemes as discussed in Mackey (2009) are appropriate within our framework.

To optimize (4), first notice the following:

Property 2. *If $P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ are convex, then* (4) *is a bi-concave problem in* $\mathbf{u}$ *and* $\mathbf{v}$.

Recall that bi-concavity implies that (4) is concave in $\mathbf{u}$ with $\mathbf{v}$ fixed as well as the converse. This observation leads to a simple alternating optimization strategy: iteratively solve (4) for $\mathbf{u}$ with $\mathbf{v}$ fixed and then for $\mathbf{v}$ with $\mathbf{u}$ fixed. This alternating scheme can also be employed when $P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ are non-convex. We use a key innovation, namely that the solution to our problem can be found by solving a penalized regression problem and re-scaling, to derive a simple and tractable proximal gradient ascent scheme (Nesterov, 2005; Beck and Teboulle, 2009). Our approach is summarized in Algorithm 1. Note that $\text{prox}_P(\mathbf{y}, \lambda) = \text{argmin}_{\mathbf{x}}\{\frac{1}{2}||\mathbf{x} - \mathbf{y}||_2^2 + \lambda P(\mathbf{x})\}$ is the proximal operator of $P()$ (Beck and Teboulle, 2009), $\lambda_{max}(\mathbf{A})$ denotes the largest eigenvalue of $\mathbf{A}$, and $||\mathbf{x}||_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A}\, \mathbf{x}}$ denotes the norm induced by $\mathbf{A}$.

Algorithm 1 Rank-one SFPCA Algorithm

1. Initialize $\mathbf{u}$ and $\mathbf{v}$ to that of the rank-1 SVD of $\mathbf{X}$. Set $\mathbf{S}_{\mathbf{u}} = \mathbf{I} + \alpha_{\mathbf{u}} \mathbf{\Omega}_{\mathbf{u}}$ and $L_{\mathbf{u}} = \lambda_{max}(\mathbf{S}_{\mathbf{u}})$; set $\mathbf{S}_{\mathbf{v}} = \mathbf{I} + \alpha_{\mathbf{v}} \mathbf{\Omega}_{\mathbf{v}}$ and $L_{\mathbf{v}} = \lambda_{max}(\mathbf{S}_{\mathbf{v}})$.
2. Repeat until convergence:
   (a) Estimate $\hat{\mathbf{u}}$. Repeat until convergence: $\mathbf{u}^{(t+1)} = \text{prox}_{P_{\mathbf{u}}}\left(\mathbf{u}^{(t)} + \frac{1}{L_{\mathbf{u}}}(\mathbf{X}\,\mathbf{v}^* - \mathbf{S}_{\mathbf{u}}\,\mathbf{u}^{(t)}), \frac{\lambda_{\mathbf{u}}}{L_{\mathbf{u}}}\right)$.
   (b) Set $\mathbf{u}^* = \begin{cases} \hat{\mathbf{u}}/||\hat{\mathbf{u}}||_{\mathbf{S}_{\mathbf{u}}} & ||\hat{\mathbf{u}}||_{\mathbf{S}_{\mathbf{u}}} > 0 \\ 0 & \text{otherwise.} \end{cases}$
   (c) Estimate $\hat{\mathbf{v}}$. Repeat until convergence: $\mathbf{v}^{(t+1)} = \text{prox}_{P_{\mathbf{v}}}\left(\mathbf{v}^{(t)} + \frac{1}{L_{\mathbf{v}}}(\mathbf{X}^T\,\mathbf{u}^* - \mathbf{S}_{\mathbf{v}}\,\mathbf{v}^{(t)}), \frac{\lambda_{\mathbf{v}}}{L_{\mathbf{v}}}\right)$.
   (d) Set $\mathbf{v}^* = \begin{cases} \hat{\mathbf{v}}/||\hat{\mathbf{v}}||_{\mathbf{S}_{\mathbf{v}}} & ||\hat{\mathbf{v}}||_{\mathbf{S}_{\mathbf{v}}} > 0 \\ 0 & \text{otherwise.} \end{cases}$
3. Return $\mathbf{u} = \mathbf{u}^* / ||\,\mathbf{u}^*\,||_2$, $\mathbf{v} = \mathbf{v}^* / ||\,\mathbf{v}^*\,||_2$, and $d = \mathbf{u}^T\,\mathbf{X}\,\mathbf{v}$.

Our algorithm enjoys strong convergence guarantees:

Theorem 2. *Assume A1-A3 and consider Algorithm 1:*

(i) Step 2 (a) monotonically converges to a global or local solution of

$$\underset{\mathbf{u}}{\text{minimize}}\, \frac{1}{2}||\,\mathbf{X}\,\mathbf{v}^* - \mathbf{u}\,||_2^2 + \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}) + \frac{\alpha_{\mathbf{u}}}{2}\,\mathbf{u}^T\,\mathbf{\Omega}_u\,\mathbf{u}, \tag{6}$$

for $P_{\mathbf{u}}$ convex or non-convex respectively. Step 2 (c) converges to a global or local solution analogously for $\mathbf{v}$.

(ii) If $P_{\mathbf{u}}()$ is convex, and letting $f(\mathbf{u})$ be the objective function of (6)*, Step 2 (a) converges at rate $f(\mathbf{u}^{(t)}) - f(\hat{\mathbf{u}}) \leq \frac{L_{\mathbf{u}}}{2t}||\,\mathbf{u}^{(0)} - \hat{\mathbf{u}}||_2^2$. Step 2 (c) converges at the analogous rate.*

(iii) Step 2 (b) is a global or local solution to (4) *for* $\mathbf{u}$ *with* $\mathbf{v}$ *fixed at* $\mathbf{v}^*$ *for $P_{\mathbf{u}}$ convex or non-convex respectively. Step 2 (d) is analogously a global or local solution to* (4) *for* $\mathbf{v}$ *with* $\mathbf{u}$ *fixed at* $\mathbf{u}^*$.

(iv) The SFPCA Algorithm monotonically converges to a stationary point of (4) *for both $P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ convex; otherwise it converges to a local optimum.*

Theorem 2 illustrates that this simple algorithmic scheme, alternating solving for $\mathbf{u}$ and $\mathbf{v}$ by fitting penalized regression problems followed by re-scaling, converges to a stationary point of (4) for convex penalties. Note that this is a stronger statement than any analogous results derived in the current literature on regression formulations of SPCA. The proximal operators employed in Steps 2 (a) and (c) have simple, closed form solutions for many of the popular sparsity penalties (Gong et al., 2013) and are typically some form of thresholding function; the proximal operator of the ℓ_1 penalty, for example, is the soft-thresholding function, $\text{prox}(x, \lambda) = \text{sign}(x)(|x| - \lambda)_+$. Also, notice that our algorithm is tractable in big-data settings with large n and/or large p as no matrix inversions or repeated eigenvalue decompositions are needed. Instead, our approach utilizes a simple iterative thresholding scheme with a total order of $O(n^2 + p^2)$ computations for Step 2. Showing the equivalence of the SFPCA problem, (4), with respect to $\mathbf{u}$ to the penalized regression problem (6) with a re-scaled solution is the key innovation that permits us to develop a fast and tractable algorithmic approach. A couple of further remarks and results are warranted.

Remark 3. *We present a proximal gradient scheme for solving the sub-problems for* $\mathbf{u}$ *and* $\mathbf{v}$ *for simplicity. The convergence rate of these methods,* $O(1/t)$*, can be improved to* $O(1/t^2)$ *by employing fast proximal gradient methods as described in Beck and Teboulle (2009).*

Also, since our problem, (4), is non-convex, there are potentially many stationary points and the stationary point reached by our algorithm will depend on the initializations for $\mathbf{u}$ and $\mathbf{v}$:

Property 3. *With convex penalties, then given an initialization for* $\mathbf{u}$ *and* $\mathbf{v}$ *in Algorithm 1, the resulting SFPCA solution is unique.*

We recommend, then, to initialize the algorithm to the rank-one SVD (PCA) solution to assure that a good stationary point is reached.

Now consider additionally adding non-negativity constraints, (5), to our SFPCA problem:

Corollary 1. *Replace* $\text{prox}_{P_\mathbf{u}}()$ *in Step 2 (a) with the positive proximal operator,* $\text{prox}_P^+(\mathbf{y}, \lambda) = \text{argmin}_{\mathbf{x}:\mathbf{x}\geq 0}\{\frac{1}{2}||\mathbf{x} - \mathbf{y}||_2^2 + \lambda P(\mathbf{x})\}$*, and analogously for Step 2 (c). Then, Theorem 2 holds for the associated regression problem* (6) *and SFPCA problem* (5) *with non-negativity constraints.*

This positive proximal operator also has simple closed form solutions for many of the popular sparsity penalties; with ℓ_1 penalties, for example, $\text{prox}^+(y, \lambda) = (x - \lambda)_+$, the positive soft-thresholding operator. Thus, we have developed a general algorithmic strategy tractable in big-data settings that can achieve many types of regularization in PCA.

4.1 Selecting Regularization Parameters

When applying our SFPCA algorithm in practice, one must use data-driven methods to estimate the optimal regularization parameters, $\lambda_\mathbf{u}$, $\lambda_\mathbf{v}$, $\alpha_\mathbf{u}$, and $\alpha_\mathbf{v}$ controlling the amount of sparsity and smoothness. Notably, the need to estimate four separate penalty parameters may seem like a major drawback of our formulation. Indeed, searching over a four dimensional grid of regularization parameters using data-driven methods suggested in the literature such as cross-validation (Silverman, 1996; Shen and Huang, 2008; Witten et al., 2009) would be a major computational burden. Due to the close connection to penalized regression methods, however, we are able to dramatically reduce this burden by expanding on an idea introduced by Huang et al. (2009) for two-way FPCA methods: Since an alternating algorithm is employed estimating $\mathbf{u}$ with $\mathbf{v}$ fixed and vice versa, they suggest to find the optimal penalty parameter in a nested manner by finding the best penalty parameter for $\mathbf{u}$ while performing the ascent step for $\mathbf{u}$ and analogously for $\mathbf{v}$.

We adopt a similar strategy suited for estimating both the sparsity and smoothness parameters $\lambda_\mathbf{u}$ and $\alpha_\mathbf{u}$ together by embedding model selection in the alternating steps of our SFPCA algorithm. Specifically for the case of ℓ_1 sparsity penalties, notice that the update for $\mathbf{u}$ is given by solving an ℓ_1 and generalized ℓ_2 penalized regression problem and then re-scaling, a similar regression problem to the elastic net of Zou and Hastie (2005). While many model selection schemes have been employed for the elastic net, the BIC method is a simple and well-studied method as there exists a closed form estimate of the degrees of freedom. We adapt this for our generalized ℓ_2 penalty:

Proposition 2. *An estimate of the degrees of freedom for* (6) *with* ℓ_1 *penalties is given by* $\hat{df}(\hat{\mathbf{u}}) = \text{tr}[\mathbf{I}_{|\mathcal{A}|} - \alpha_\mathbf{u}\,\mathbf{\Omega}_\mathbf{u}(\mathcal{A}, \mathcal{A})]$ *where* $\mathcal{A}$ *denotes the indices of the estimated non-zeros in* $\hat{\mathbf{u}}$*. The BIC selection criterion is given by* $BIC(\hat{\mathbf{u}}) = \log[\frac{1}{2n}||\,\mathbf{X}\,\mathbf{v}^* - \hat{\mathbf{u}}||_2^2] + \frac{1}{n}\log(n)\hat{df}(\hat{\mathbf{u}})$*, with the criterion for* $\mathbf{v}$ *defined analogously.*

Similar degrees of freedom estimates can be calculated for other order-one penalties (Tibshirani and Taylor, 2012). Thus, instead of performing a grid search over all four parameter values, we can search over those for $\mathbf{u}$ and $\mathbf{v}$ separately, akin to other two-way regularized PCA methods (Huang et al., 2009; Lee et al., 2010). Furthermore, warm starts can be used for the sparsity parameters with the proximal gradient algorithm to swiftly fit the algorithm over a sequence of parameters. Now, one potential criticism of this nested approach is that the penalty parameters selected at each step are not guaranteed to converge. Thus, we suggest to perform nested model selection for a small number of steps and then run Algorithm 1 at the selected penalty parameters to ensure convergence to a stationary point. In our experience, the penalty parameters from nested selection tend to stabilize quickly. As shown in the subsequent numerical studies, this scheme performs well, selecting data-driven flexible combinations of smoothness and sparsity in a tractable computational manner.

Also in practice, one needs to estimate the rank of the model to greedily extract. Several methods have been suggested in the PCA and regularized PCA literature (Buja and Eyuboglu, 1992; Troyanskaya et al., 2001; Owen and Perry, 2009) that are also appropriate in our framework.

5 Results

5.1 Simulation Study

We evaluate the comparative performance of our SFPCA method on simulated data sets. Data is simulated according to the low rank model: $\mathbf{X}_{n\times p} = \sum_{k=1}^{K} d_k \, \mathbf{u}_k \, \mathbf{v}_k^T + \epsilon$, where $\epsilon_{ij} \overset{iid}{\sim} N(0,1)$, $\mathbf{v}_k$ is the fixed signal, a smooth and sparse unit vector of length p, $\mathbf{u}_k$ is a random unit vector of length n, and d_k modulates the signal-to-noise ratio. The fixed signal vectors, $\mathbf{v}$, are fixed at length $p = 200$ with sparsity and smoothness in the nature of a time series: $\mathbf{v}_1$ is a sinusoidal curve with phase $t = 60$ and 31% sparsity; $\mathbf{v}_2$ is a Gaussian-modulated sinusoidal pulse with frequency $t = 7$ and 25% sparsity; and, $\mathbf{v}_3$ is a sinusoidal curve with phase $t = 30$ and sparsity 31%. Figures illustrating these true signal factors are given in the top left panel of Figure 2. The left PCs, $\mathbf{u}$, are generated as the left singular vectors of a random Gaussian matrix with identity covariance. The scale-factor which modulates the signal-to-noise ratio is taken as a function of the sample size: $d_1 = n/4$, $d_2 = n/5$ and $d_3 = n/6$.

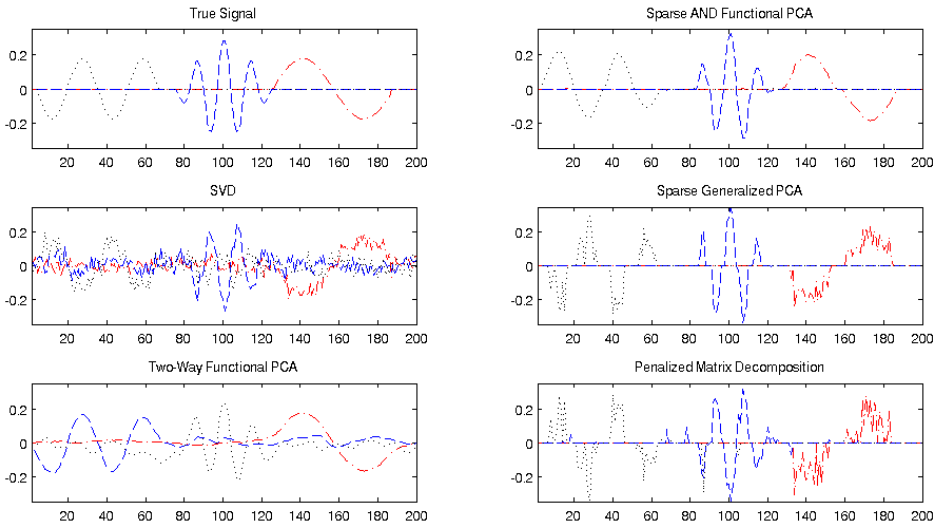


Figure 2: Examples of true factors, $\mathbf{v}_1$ (red, dotted-dashed), $\mathbf{v}_2$ (blue, dashed) and $\mathbf{v}_3$ (black, dotted), and those estimated by various methods for the simulation study.

We compare our method to other recently proposed approaches for regularized PCA: Two-way FPCA (TWFPCA) of Huang et al. (2009) which is the two-way generalization of the FPCA method of Silverman (1996); Huang et al. (2008); Sparse SVD (SSVD) of Lee et al. (2010) which is a two-way generalization of the SPCA method of Shen and Huang (2008); Penalized Matrix Decomposition (PMD) of Witten et al. (2009); and Sparse Generalized PCA

Table 1: Simulation results averaged over 50 replicates in terms of 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). Best performing methods are boldfaced.

			TWFPCA	SSVD	PMD	SGPCA ($\sigma = 1$)	SGPCA ($\sigma = 5$)	SFPCA
$n = 100$	$\mathbf{v}_1$	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)
	$\mathbf{v}_2$	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)
	$\mathbf{v}_3$	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)
$n = 300$	$\mathbf{v}_1$	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)
	$\mathbf{v}_2$	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)
	$\mathbf{v}_3$	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)

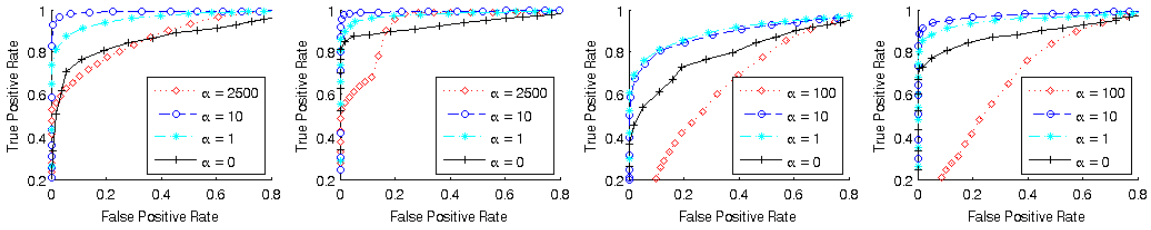


Figure 3: Mean ROC curves computed by varying the sparsity parameter, λ, for SFPCA methods at differing levels of smoothness, α. From left to right, results are given for rank-1 models with signal $\mathbf{v}_1$ and $n = 50$, $\mathbf{v}_1$ and $n = 100$, $\mathbf{v}_2$ and $n = 50$, and $\mathbf{v}_2$ and $n = 100$; the SNR for all scenarios is determined by setting $d = n/2$.

(SGPCA) of Allen et al. (2011). For the latter, generalizing operators must be specified and $e^{-d_{ij}^2/\sigma}$ was used as suggested in Allen and Maletić-Savatić (2011) for Chebychev distances between time points, d_{ij}, and parameter σ. Smoothing matrices must be specified for TWFPCA and our SFPCA method and were taken as the squared second differences matrix. All sparse methods, including our own, were implemented with ℓ_1 sparsity penalties. All methods use regularization parameter selection schemes as described in the corresponding papers and were implemented in Matlab with the exception of the PMD which was implemented via the corresponding `R` package `PMA` (Witten et al., 2009). Our SFPCA method was also coded in Matlab and utilized the BIC to perform data-driven regularization parameter selection as described previously.

In Table 1, results on signal recovery and feature selection are presented. The true positive rate (TP) and false positive rate (FP) are reported for recovering the support of $\mathbf{v}$ as well as two measures of smoothness, the relative angle and relative squared error. The relative angle (r$\angle$) is given by r$\angle = (1 - |\hat{\mathbf{v}}^T \mathbf{v}^*|)/(1 - |\mathbf{v}_{(SVD)}^T \mathbf{v}^*|)$ where $\mathbf{v}^*$ is the true signal and $\mathbf{v}_{(SVD)}$ is that recovered by the SVD. As the cosine of the angle between two unit vectors is the inner product, this relative angle measures how close $\hat{\mathbf{v}}$ is to $\mathbf{v}^*$ relative to that of the SVD; smaller values then indicate better signal recovery. The relative squared error (rSE) is given by rSE $= ||\mathbf{X}^* - \hat{\mathbf{X}}||_F^2 / ||\mathbf{X}^* - \mathbf{X}^{(SVD)}||_F^2$ with smaller values indicating better low-rank signal recovery than the SVD. Results in Table 1 indicate that SFPCA performs better than all competing methods and the SVD in terms of both feature selection AND smooth signal recovery. Example plots of the recovered PC loadings by each of the methods are shown in Figure 2.

The results of our numerical study reveal that SFPCA outperforms existing methods for SPCA in terms of feature selection. To further investigate this phenomena, we compute receiver-operator-characteristic (ROC) curves, Figure 3, for recovering $\mathbf{v}_1$ and $\mathbf{v}_2$ in the above simulation by varying the sparsity parameter, λ, under differing amounts of smoothness α. Results indicate that encouraging some smoothness does indeed lead to better feature selection results

than existing SPCA approaches where $\alpha = 0$. Overall, our numerical simulations have shown that SFPCA leads to not only improved smooth signal recovery of the underlying low-rank structure, but it also improved feature selection. An additional two-way simulation study is given in Appendix B.

6 Case Study: EEG Data

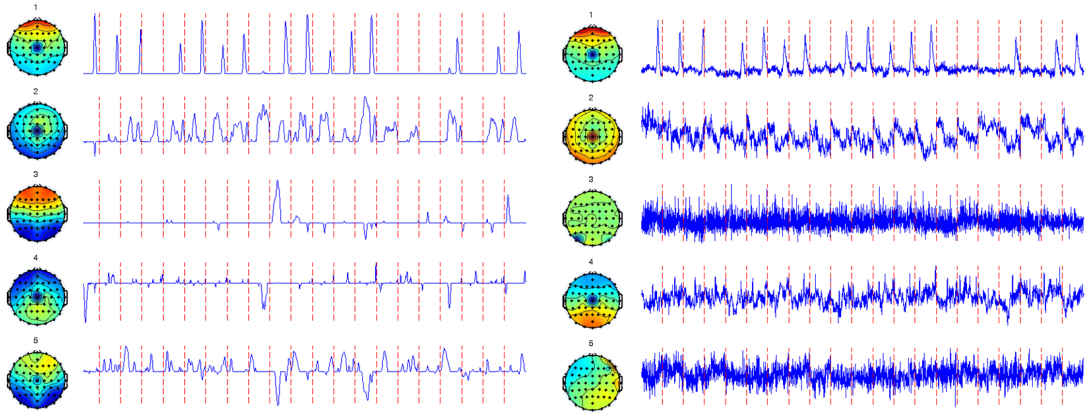


Figure 4: First five spatial and temporal SFPCA components (left) and ICA components (right) for the EEG case study.

We test our method on an electroencephalography (EEG) case study on genetic predisposition to alcoholism. Publicly available data was downloaded from `http://archive.ics.uci.edu/ml/datasets/EEG+Database` which was sampled from 64 channels at 256Hz; the first alcoholic subject over epochs relating to non-matching stimuli was examined. The data consists of $n = 57$ EEG channels with corresponding scalp locations and $p = 5376$ time points corresponding to 21 epochs of 256 time points each. 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 visualize the data (Makeig et al., 1996; Hyvärinen and Oja, 2000). Our SFPCA method was applied to this data set with the spatial smoothing matrix, $\mathbf{\Omega_u}$, specified as the weighted squared second differences matrix using spherical distances between the channel locations and the temporal smoothing matrix, $\mathbf{\Omega_v}$, taken as the squared second differences matrix. The nested BIC method was used to select all four penalty parameters, $\lambda_{\mathbf{u}}$,$\lambda_{\mathbf{v}}$, $\alpha_{\mathbf{u}}$, and $\alpha_{\mathbf{v}}$. We also apply ICA via the FastICA method (Hyvärinen and Oja, 2000) as well as PCA, two-way FPCA (Huang et al., 2009), two-way Sparse PCA (Witten et al., 2009), and two-way Sparse Generalized PCA (Allen et al., 2011). Results for the latter four methods are shown in Appendix C.

Results for the first five estimated spatial and temporal components of the EEG data for SFPCA and ICA are shown in Figure 6. Our algorithm is tractable for high-dimensional data, converging in 3.28 seconds for software programmed entirely in Matlab. Also, because our optimization framework permits combinations of smoothness and sparsity, our method is conducive to data-driven selection of the type of regularization best to model the data. For this EEG example, our method selected $\lambda_{\mathbf{u}} = 0$ for each of the first five components, meaning that sparsity in the EEG channels was not needed. A range of smoothing parameters, $\alpha_{\mathbf{u}} = [10, 12]$ and $\alpha_{\mathbf{v}} = [0.5, 10]$ as well as temporal sparsity parameters, $\lambda_{\mathbf{v}} = [1, 2.5]$, were selected for the first five components, indicating that our method is able to flexibly choose the optimal amount of smoothness and sparsity for recovering major patterns in the data. Compared to ICA, our method extracts the same major patterns and source localizations. But, due to smoothing and feature selection, the resulting epoch-time series are much more interpretable. (Figure 6 illustrates the onset of each new epoch in the time series by a red vertical dotted line.) The improvements are clearly seen by examining the first two SFPCA and ICA components which have the same peaks in the temporal activation patterns; but because the noise level of our model is reduced by sparsity and smoothing, these peaks are much more apparent with our SFPCA method. The resulting SFPCA components are also much more interpretable compared to other PCA and regularized PCA methods as presented in Appendix C.

7 Discussion

In this paper, we have introduced a regularized PCA method to achieve sparsity and/or smoothness of the row and/or column factors. Key contributions include developing a unified optimization problem that (i) generalizes much of the existing regularized PCA literature, (ii) has the flexibility to determine the types and amounts of regularization in a data-driven manner, (iii) can accommodate a variety of sparsity penalties and/or non-negativity constraints, (iv) enjoys desirable numerical properties, (v) is conducive to a simple, scalable algorithm to optimize the model with notable convergence guarantees, and (vi) has strong demonstrated numerical results; additionally, we have (vii) provided results on the types of general sparse penalties leading to convergent regularized PC solutions.

There are many avenues of future statistical research related to our work. The focus of this paper has been on formulating and studying an optimization model that unifies regularization in PCA; by doing so, we establish a foundation upon which to study and perhaps unify other theoretical aspects of regularized PCA methods such as consistency in high-dimensional settings. As special cases of our model have been shown to lead to consistent PC factors (Silverman, 1996; Shen et al., 2012), we conjecture that our method also leads to consistent estimates, an interesting topic for future research. The important considerations we propose for formulating an optimization problem as well as the significant contribution of generally defining the types of sparse penalties that lead to convergent solutions could also be explored in the context of other dimension reduction techniques such as CCA, LDA, PLS, kernel PCA, and spectral clustering. Finally, our approach could also be extended to multi-way or tensor data following the framework introduced in Allen (2012).

Sparse PCA or Functional PCA have been widely used in a variety of applied fields. As our flexible method enjoys improved interpretability and feature selection over these existing regularized PCA methods, there is potential for broad application of our approach to structured big-data in areas such as neuroimaging, genomics, proteomics, epidemiology, climate studies, spatial data, time series, imaging, compressed sensing, and engineering. `Matlab` functions implementing SFPCA are available from `http://www.stat.rice.edu/~gallen/software.html`.

Acknowledgments

Thanks to Yue Hu for help with the EEG data. This work is supported by nsf-dms 1209017.

References

Allen, G. I. (2012). Sparse higher-order principal components analysis. In *AISTATS*, Volume 15.

Allen, G. I., L. Grosenick, and J. Taylor (2011). A generalized least squares matrix decomposition. Rice University Technical Report No. TR2011-03.

Allen, G. I. and M. Maletić-Savatić (2011). Sparse non-negative generalized pca with applications to metabolomics. *Bioinformatics 27*(21), 3029–3035.

Amini, A. and M. Wainwright (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics 37*(5B), 2877–2921.

Bach, F., R. Jenatton, J. Mairal, and G. Obozinski (2012). Structured sparsity through convex optimization. *Statistical Science 27*(4), 450–468.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Bertsekas, D. P., A. Nedić, and A. E. Ozdaglar (2003). *Convex analysis and optimization*. Athena Scientific Belmont.

Bühlmann, P. L., S. A. van de Geer, and S. Van de Geer (2011). *Statistics for high-dimensional data*. Springer.

Buja, A. and N. Eyuboglu (1992). Remarks on parallel analysis. *Multivariate Behavioral Research 27*(4), 509–540.

Cai, T. T., Z. Ma, and Y. Wu (2012). Sparse pca: Optimal rates and adaptive estimation. *arXiv preprint arXiv:1211.1309*.

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

d'Aspremont, A., L. El Ghaoui, M. I. Jordan, and G. Lanckriet (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM review 49*(3), 434–448.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Gong, P., C. Zhang, Z. Lu, J. Huang, and J. Ye (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *arXiv preprint arXiv:1303.4434*.

Hebiri, M. and S. van de Geer (2011). The smooth-lasso and other l1+ l2-penalized methods. *Electronic Journal of Statistics 5*, 1184–1226.

Huang, J., H. Shen, and A. Buja (2008). Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics 2*, 678–695.

Huang, J., H. Shen, and A. Buja (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association 104*(488), 1609–1620.

Hyvärinen, A. and E. Oja (2000). Independent component analysis: algorithms and applications. *Neural networks 13*(4), 411–430.

Jenatton, R., G. Obozinski, and F. Bach (2009). Structured sparse principal component analysis. *arXiv preprint arXiv:0909.1440*.

Johnstone, I. and A. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104*(486), 682–693.

Jolliffe, I., N. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

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

Jung, S. and J. Marron (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics 37*(6B), 4104–4130.

Lee, M., H. Shen, J. Huang, and J. Marron (2010). Biclustering via Sparse Singular Value Decomposition. *Biometrics 66*(4), 1087–1095.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41*(2), 772–801.

Mackey, L. (2009). Deflation methods for sparse pca. *Advances in neural information processing systems 21*, 1017–1024.

Makeig, S., A. J. Bell, T. Jung, T. J. Sejnowski, et al. (1996). Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, 145–151.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming 103*(1), 127–152.

Owen, A. and P. Perry (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Annals 3*(2), 564–594.

Ramsay, J. (2006). *Functional data analysis*. Wiley Online Library.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B*, 233–243.

Rockafellar, R. T. (1997). *Convex analysis*, Volume 28. Princeton university press.

Shen, D., H. Shen, and J. Marron (2012). Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis*.

Shen, H. and J. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis 99*(6), 1015–1034.

Silverman, B. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics 24*(1), 1–24.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Tibshirani, R. J. and J. Taylor (2012). Degrees of freedom in lasso problems. *The Annals of Statistics 40*(2), 1198–1232.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics 17*(6), 520.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications 109*(3), 475–494.

Vu, V. Q. and J. Lei (2012). Minimax rates of estimation for sparse pca in high dimensions. *arXiv preprint arXiv:1202.0786*.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Yang, D., Z. Ma, and A. Buja (2011). A sparse svd method for high-dimensional data. *arXiv preprint arXiv:1112.2433*.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zass, R. and A. Shashua (2007). Nonnegative sparse pca. *Advances in Neural Information Processing Systems 19*, 1561.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.

A Proofs

We begin with a few results that will be useful in proving the two major theorems in this paper.

Proposition 3. *An optimal pair* $(\mathbf{u}, \mathbf{v})$ *of the SFPCA problem,* (4)*, necessarily satisfies the following Karush-Kuhn-Tucker (KKT) conditions:*

$$\mathbf{X}\,\mathbf{v} - \lambda_{\mathbf{u}}\Gamma_{\mathbf{u}}(\mathbf{u}) - 2\gamma_{\mathbf{u}}\,\mathbf{S}_{\mathbf{u}}\,\mathbf{u} = 0 \tag{7}$$

$$\gamma_{\mathbf{u}}(\mathbf{u}^T\,\mathbf{S}_{\mathbf{u}}\,\mathbf{u} - 1) = 0 \tag{8}$$

$$\mathbf{X}^T\,\mathbf{u} - \lambda_{\mathbf{v}}\Gamma_{\mathbf{v}}(\mathbf{v}) - 2\gamma_{\mathbf{v}}\,\mathbf{S}_{\mathbf{v}}\,\mathbf{v} = 0 \tag{9}$$

$$\gamma_{\mathbf{v}}(\mathbf{v}^T\,\mathbf{S}_{\mathbf{v}}\,\mathbf{v} - 1) = 0 \tag{10}$$

$$\gamma_{\mathbf{u}}, \gamma_{\mathbf{v}} \geq 0 \tag{11}$$

$$\mathbf{u}^T\,\mathbf{S}_{\mathbf{u}}\,\mathbf{u}, \mathbf{v}^T\,\mathbf{S}_{\mathbf{v}}\,\mathbf{v} \leq 1. \tag{12}$$

Here, the γ*'s are dual variables associated with the inequality constraints of* (4) *and the* $\Gamma(\mathbf{x})$*'s are sub-gradients of the associated* $P(\mathbf{x})$*, defined as the set of functions that satisfy:*

$$P(\mathbf{y}) \geq P(\mathbf{x}) + \Gamma(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \;\; \forall \;\; \mathbf{y} \in \mathbb{R}\,. \tag{13}$$

Proof. Even for non-convex optimization problems, zero duality gap can hold implying that the KKT conditions are necessary for optimality under certain constraint qualifications (Bertsekas et al., 2003). For (4), $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^p$ are convex sets, each inequality constraint is convex in $\mathbf{u}$ or $\mathbf{v}$ and jointly convex in both, and there trivially exists a strictly feasible pair, $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$: $\tilde{\mathbf{u}}^T\,\mathbf{S}_{\mathbf{u}}\,\tilde{\mathbf{u}} < 1$ and $\tilde{\mathbf{v}}^T\,\mathbf{S}_{\mathbf{v}}\,\tilde{\mathbf{v}} < 1$ for $\mathbf{S} \succ 0$ and finite $\lambda_{max}(\mathbf{S}) < M$. (4) then satisfies Slater's constraint qualification for non-convex problems (see Bertsekas et al., 2003, Ch. 5), and thus, the KKT conditions are necessary for optimality. □

Lemma 1. *For penalty functions* $P()$ *satisfying assumption A2, if* $\Gamma(\mathbf{x})$ *is a subgradient of* $P(\mathbf{x})$*, then* $\Gamma(c\mathbf{x}) \;\; \forall\, c > 0$ *is also a subgradient of* $P(\mathbf{x})$*.*

Proof. Subgradients are defined as any function that satisfies (13). Substitute $\tilde{\mathbf{x}} = c\mathbf{x}$ in (13):

$$\begin{aligned}
&P(\mathbf{y}) - \Gamma(c\mathbf{x})^T\mathbf{y} \geq P(c\mathbf{x}) - \Gamma(c\mathbf{x})^T(c\mathbf{x}) \\
&\frac{1}{c}P(\mathbf{y}) - \frac{1}{c}\Gamma(c\mathbf{x})^T\mathbf{y} \geq P(\mathbf{x}) - \Gamma(c\mathbf{x})^T\mathbf{x} \quad \because P(c\mathbf{x}) = cP(\mathbf{x}) \\
&P(\tilde{\mathbf{y}}) - \Gamma(c\mathbf{x})^T\tilde{\mathbf{y}} \geq P(\mathbf{x}) - \Gamma(c\mathbf{x})^T\mathbf{x} \;\; \forall \;\; \tilde{\mathbf{y}} = \frac{1}{c}\mathbf{y} \in \mathbb{R}\,.
\end{aligned}$$

Since the last inequality holds for all $\tilde{\mathbf{y}}$ in the domain of the penalties, this satisfies the definition of a subgradient (13). Hence, $\Gamma(c\mathbf{x})$ is also a subgradient of $P(\mathbf{x})$. □

Proposition 4 (Part (iii) of Theorem 2)**.** *For penalty function* $P()$ *satisfying assumption A2, the solution,* $\mathbf{u}^*$*, to the following optimization problem:*

$$\underset{\mathbf{u}}{\text{maximize}} \;\; \mathbf{u}^T\,\mathbf{X}\,\mathbf{v} - \lambda P(\mathbf{u}) \;\; \text{subject to} \;\; \mathbf{u}^T\,\mathbf{S}\,\mathbf{u} \leq 1, \tag{14}$$

where $\mathbf{S} = \mathbf{I} + \alpha\,\mathbf{\Omega}$*, is given by:*

$$\hat{\mathbf{u}} = \underset{\mathbf{u}}{\text{argmin}} \left\{ \frac{1}{2}||\,\mathbf{X}\,\mathbf{v} - \mathbf{u}\,||_2^2 + \lambda P(\mathbf{u}) + \frac{\alpha}{2}\,\mathbf{u}^T\,\mathbf{\Omega}\,\mathbf{u} \right\} \tag{15}$$

$$\mathbf{u}^* = \begin{cases} \hat{\mathbf{u}}/||\hat{\mathbf{u}}||_{\mathbf{S}} & ||\hat{\mathbf{u}}||_{\mathbf{S}} > 0 \\ 0 & \textit{otherwise.} \end{cases} \tag{16}$$

$\mathbf{u}^*$ *is the global solution if* $P()$ *is convex and a local solution otherwise.*

Proof. (Note that this result and proof is an extension of a result in Allen et al. (2011).) From Proposition 3, we see that the KKT conditions are necessary for optimality even when $P()$ is non-convex. The first-order KKT conditions for (14) are given by (7) (re-stated for completeness here):

$$2\gamma\,\mathbf{S}\,\mathbf{u} - \mathbf{X}\,\mathbf{v} + \lambda\Gamma(\mathbf{u}) = 0 \tag{17}$$

and the complementary slackness by (8). On the other hand, the first-order KKT condition for (15) is given by

$$\mathbf{S}\,\mathbf{u} - \mathbf{X}\,\mathbf{v} + \lambda\Gamma(\mathbf{u}) = 0. \tag{18}$$

Note that (18) is nearly equivalent to (17), the first order condition of (4), except for the factor of 2γ on the first term. Now, let $\hat{\mathbf{u}}$ be the solution to (6), that is $\hat{\mathbf{u}}$ solves (18). For some constant $c > 0$, let $\tilde{\mathbf{u}} = c\hat{\mathbf{u}}$. We can re-write (18) as:

$$\begin{aligned} \frac{1}{c}\,\mathbf{S}\,\tilde{\mathbf{u}} - \mathbf{X}\,\mathbf{v} + \lambda\Gamma(\frac{1}{c}\tilde{\mathbf{u}}) = 0, \\ \frac{1}{c}\,\mathbf{S}\,\tilde{\mathbf{u}} - \mathbf{X}\,\mathbf{v} + \lambda\Gamma(\tilde{\mathbf{u}}) = 0, \;\; \because \text{Lemma } 1. \end{aligned}$$

Now, letting $c = 1/2\gamma$, we see that (18) is equivalent to (17). Next, consider the pair $(\mathbf{u}^* = \hat{\mathbf{u}}/||\hat{\mathbf{u}}||_{\mathbf{S}}, \gamma^* = ||\hat{\mathbf{u}}||_{\mathbf{S}}/2)$. Then, notice that $(\mathbf{u}^*, \gamma^*)$ satisfy both (17),(8) and are hence a solution for (4) with respect to $\mathbf{u}$. If $P()$ is convex, then Slater's condition implies that the KKT conditions are necessary and sufficient for global optimality; hence, $\mathbf{u}^*$ is the global solution in this case. □

of Theorem 1. We will prove that the solution to our SFPCA problem, (4), is equivalent to the solutions of the stated problems by equating the Karush-Kuhn-Tucker (KKT) conditions of the optimization problems. From Proposition 3, we have the the KKT conditions for (4) are necessary for optimality; the same argument can be used to show that the KKT conditions are necessary for optimality of the SPCA, two-way SPCA, and two-way FPCA problems of Shen and Huang (2008); Witten et al. (2009); Huang et al. (2009) respectively. Then for these cases to show equivalence with our SFPCA problem, it suffices to show that the KKT conditions of the underlying optimization problems are equivalent. Now, the PCA and FPCA (Silverman, 1996) problems are eigenvalue and generalized eigenvalue problems respectively; we will show that our SFPCA optimization problem is equivalent to these.

Proof of part (i): Consider the PCA/SVD optimization problem: $\text{minimize}_{\mathbf{u},\mathbf{v}} \; ||\,\mathbf{X} - d\,\mathbf{u}\,\mathbf{v}^T\,||_F^2$ subject to $\mathbf{u}^T\,\mathbf{u} = 1 \;\&\; \mathbf{v}^T\,\mathbf{v} = 1$. Multiplying out the Frobenius norm in the objective, we have $\text{tr}(\mathbf{X}^T\,\mathbf{X}) - d\,\mathbf{u}^T\,\mathbf{X}\,\mathbf{v} + d^2\,\mathbf{u}^T\,\mathbf{u}\,\mathbf{v}^T\,\mathbf{v}$. The first term does not depend on the variables, and the last term does not depend on $\mathbf{u}$ and $\mathbf{v}$ because of the constraints. Thus, the objectives of (4) when $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$ and the SVD problem are equivalent. Now, to show that at the solution the constraints of (4), $\mathbf{u}^T\,\mathbf{u} \leq 1$ and $\mathbf{v}^T\,\mathbf{v} \leq 1$ must be tight, consider the case when $\mathbf{u}^T\,\mathbf{u} < 1$. Then, we can define a constant $c > 0$ such that $(c\,\mathbf{u})^T(c\,\mathbf{u}) = 1$. But, then the objective, $(c\,\mathbf{u})^T\,\mathbf{X}\,\mathbf{v} < \mathbf{u}^T\,\mathbf{X}\,\mathbf{v}$, which contradicts the fact that $\mathbf{u}$ is an optimal point. Thus, we have shown that both the objectives and constraint regions of (4) when all regularization parameters are zero are equivalent to that of the SVD problem.

Proof of part (ii): The proof of this result follows closely from the argument used to prove Proposition 4. The SPCA problem considered by Shen and Huang (2008) is given by:

$$\underset{\mathbf{u},\mathbf{v}}{\text{minimize}} \;\; ||\,\mathbf{X} - \mathbf{u}\,\mathbf{v}^T\,||_F^2 + \lambda P(\mathbf{v}) \;\; \text{subject to} \;\; \mathbf{u}^T\,\mathbf{u} = 1.$$

Consider the first order KKT conditions of this problem for $\mathbf{v}$:

$$\mathbf{X}^T\,\mathbf{u} - \mathbf{v} - \lambda\Gamma(\mathbf{v}) = 0. \tag{19}$$

The KKT conditions for optimality of $\mathbf{v}$ in our SFPCA problem when $\lambda_{\mathbf{u}}, \alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$ are given by:

$$\begin{aligned} \mathbf{X}^T\,\mathbf{u} - 2\gamma_{\mathbf{v}}\,\mathbf{v} - \lambda_{\mathbf{v}}\Gamma(\mathbf{v}) = 0 \\ \gamma_{\mathbf{v}}(\mathbf{v}^T\,\mathbf{v} - 1) = 0. \end{aligned} \tag{20}$$

Notice that (19) and (20) are nearly identical except for the factor $2\gamma_{\mathbf{v}}$ multiplying the second term of the latter. Now, if we let $\hat{\mathbf{v}}$ be the solution to (19) and define $\tilde{\mathbf{v}} = c\hat{\mathbf{v}}$ for any $c > 0$, then (19) can be written as

$$\mathbf{X}^T\,\mathbf{u} - \frac{1}{c}\tilde{\mathbf{v}} - \lambda\Gamma(\tilde{\mathbf{v}}) = 0 \;\; \because \;\; Lemma\;1.$$

If we let $c = 1/2\gamma_{\mathbf{v}}$, then (19) and (20) are identical. Therefore, our SFPCA problem with $\lambda_{\mathbf{v}}$ active is a scaled version of the SPCA solution of Shen and Huang (2008).

Proof of part (iii): When $\alpha_{\mathbf{u}}, \alpha_{\mathbf{v}} = 0$, (4) has the form of the optimization problem for Sparse Generalized PCA of Allen et al. (2011) with generalizing operators, $\mathbf{Q} = \mathbf{I}$ and $\mathbf{R} = \mathbf{I}$. Also note that this problem is the Lagrange form of that in Witten et al. (2009) and closely related to that of Lee et al. (2010) as discussed in Allen et al. (2011).

Proof of part (iv): Consider the FPCA problem proposed by Silverman (1996) which was shown to be equivalent to a penalized regression problem of Huang et al. (2008):

$$\underset{\mathbf{v}}{\text{maximize}} \;\; \mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} \;\; \text{subject to} \;\; \mathbf{v}^T \mathbf{S}_{\mathbf{v}} \mathbf{v} = 1. \tag{21}$$

Now, consider (4) with only $\alpha_{\mathbf{v}}$ active:

$$\underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \;\; \mathbf{u}^T \mathbf{X} \mathbf{v} \;\; \text{subject to} \;\; \mathbf{v}^T \mathbf{S}_{\mathbf{v}} \mathbf{v} \leq 1 \;\&\; \mathbf{u}^T \mathbf{u} \leq 1. \tag{22}$$

Solving the above problem for $\mathbf{u}$ it is easy to see that the solution is $\mathbf{u} = \mathbf{X}\mathbf{v} / \sqrt{\mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v}}$. Plugging this into the above problem we arrive at:

$$\underset{\mathbf{v}}{\text{maximize}} \;\; \sqrt{\mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v}} \;\; \text{subject to} \;\; \mathbf{v}^T \mathbf{S}_{\mathbf{v}} \mathbf{v} \leq 1. \tag{23}$$

The argument used for tightness of the constraints in part (i) also reveal that the constraint must hold with equality for all $\mathbf{S}_{\mathbf{v}} \succ 0$. Then a reformulation of our SFPCA problem is equivalent to the FPCA problem proposed by Silverman (1996).

Proof of part (v): Consider our SFPCA problem, (4), with $\alpha_{\mathbf{u}}$ and $\alpha_{\mathbf{v}}$ active:

$$\underset{\mathbf{u},\mathbf{v}}{\text{maximize}} \;\; \mathbf{u}^T \mathbf{X} \mathbf{v} \;\; \text{subject to} \;\; \mathbf{u}^T \mathbf{S}_{\mathbf{u}} \mathbf{u} \leq 1 \;\&\; \mathbf{v}^T \mathbf{S}_{\mathbf{v}} \mathbf{v} \leq 1. \tag{24}$$

Then, the KKT conditions for $\mathbf{u}$ are given by:

$$\mathbf{X}\mathbf{v} - 2\gamma_{\mathbf{u}} \mathbf{S}_{\mathbf{u}} \mathbf{u} = 0 \;\; \& \;\; \gamma_{\mathbf{u}}(\mathbf{u}^T \mathbf{S}_{\mathbf{u}} \mathbf{u} - 1) = 0,$$

and the KKT conditions for $\mathbf{v}$ are analogous. Solving the first order condition for $\mathbf{u}$ and $\mathbf{v}$ gives us that $\mathbf{u}^* \propto \mathbf{S}_{\mathbf{u}}^{-1} \mathbf{X} \mathbf{v}$ and $\mathbf{v}^* \propto \mathbf{S}_{\mathbf{v}}^{-1} \mathbf{X}^T \mathbf{u}$. Notice that these are the exact same stationary equations given by the first order conditions of the two-way FPCA problem proposed by Huang et al. (2009). Interestingly, however, the scaling factor of the two sets of stationary equations are different. For our model, the complementary slackness condition gives that $\gamma_{\mathbf{u}} = ||\,\mathbf{u}\,||_{\mathbf{S}_{\mathbf{u}}}/2$. This gives the update as $\mathbf{u}^* = \mathbf{S}_{\mathbf{u}}^{-1} \mathbf{X} \mathbf{v} \,/\, \mathbf{v}^T \mathbf{X}^T \mathbf{S}_{\mathbf{u}}^{-1} \mathbf{X} \mathbf{v}$, whereas the update in Huang et al. (2009) is given by $\mathbf{u}^* = \mathbf{S}_{\mathbf{u}}^{-1} \mathbf{X} \mathbf{v} \,/\, \mathbf{v}^T \mathbf{S}_{\mathbf{v}} \mathbf{v}$. This scaling difference, however, does not change the direction of the resulting solution vectors and in practice, is often irrelevant as all algorithms typically return vectors of norm one. □

of Property 1. The SFPCA problem, (4), does not permit scale-transforms; furthermore from Proposition 1, the scaling of the solution is unique. It is trivial to see that as with classical PCA, the solution $(\mathbf{u}, \mathbf{v})$ is invariant to all but a sign change. □

of Proposition 1. For the proof of this proposition, we refer to the KKT conditions for (4) given in Proposition 3 as well as those for (6) given in Proposition 4.

From Proposition 3, notice that $\mathbf{u}^* = \mathbf{0}$ only if $\hat{\mathbf{u}} = \mathbf{0}$. But since, (4) is a penalized regression problem, it follows from the first-order KKT condition that there exists a value, $\lambda_{\mathbf{u}}^{max}$, at which the solution, $\hat{\mathbf{u}} \equiv \mathbf{0}$; this value is determined by the subgradient of $P_{\mathbf{u}}()$, $\Gamma_{\mathbf{u}}()$, and defined as the solution to the following:

$$\lambda_{\mathbf{u}}^{max} \Gamma_{\mathbf{u}}(\mathbf{0}) = \mathbf{X}\mathbf{v}\,.$$

For example, if $P_{\mathbf{u}}(\mathbf{u}) = ||\,\mathbf{u}\,||_1$, then it is easy to see that $\lambda_{\mathbf{u}}^{max} = \max_i |(\mathbf{X}\mathbf{v})_i|$. We have established that there exists a value $\lambda_{\mathbf{u}}^{max}$ for which $\hat{\mathbf{u}}$ and hence $\mathbf{u}^* = \mathbf{0}$. A similar argument for the analogous regression problem for $\mathbf{v}$ holds, thus proving part (i).

Next, assume that $\mathbf{u}^* = \mathbf{0}$. Then, since $P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ map to the positive real line, necessarily the maximum objective value occurs at $\mathbf{v}^* = \mathbf{0}$. Thus, we have the following logical sequence: if $\lambda_{\mathbf{u}} \geq \lambda_{\mathbf{u}}^{max}$, then $\hat{\mathbf{u}} = \mathbf{0} \implies \mathbf{u}^* = \mathbf{0} \implies \mathbf{v}^* = \mathbf{0}$. Therefore, the zero solution occurs if $\lambda_{\mathbf{u}} \geq \lambda_{\mathbf{u}}^{max}$ OR $\lambda_{\mathbf{v}} \geq \lambda_{\mathbf{v}}^{max}$. Suppose we have both $\lambda_{\mathbf{u}} < \lambda_{\mathbf{u}}^{max}$ and $\lambda_{\mathbf{v}} < \lambda_{\mathbf{v}}^{max}$ and consider the solution $\mathbf{u}^*$ given in Proposition 3. This solution obviously depends on

$\alpha_{\mathbf{u}} > 0$, but also depends on the solution $\mathbf{v}^*$. As $\mathbf{v}^*$ depends on both $\lambda_{\mathbf{v}} > 0$ and $\alpha_{\mathbf{v}} > 0$, the solution for $\mathbf{u}^*$ (and hence $\mathbf{v}^*$) depends on all non-zero regularization parameters. This proves part (ii).

Consider optimizing (4) with respect to $\mathbf{u}$. Then by complementary slackness, (8), $\gamma_{\mathbf{u}} > 0$ if and only if $\mathbf{u}^T \mathbf{S}_{\mathbf{u}} \mathbf{u} = 1$. From the proof of Proposition 4, we have shown that if $\hat{\mathbf{u}}$ is the solution to (6), then $(\mathbf{u}^* = \hat{\mathbf{u}}/||\hat{\mathbf{u}}||_{\mathbf{S}_{\mathbf{u}}}, \gamma^*_{\mathbf{u}} = ||\hat{\mathbf{u}}||_{\mathbf{S}_{\mathbf{u}}}/2)$ is an optimal point of (4). Notice that $\gamma^*_{\mathbf{u}} = 0$ if and only if $\hat{\mathbf{u}} = \mathbf{0}$, which from the above argument occurs if either $\lambda_{\mathbf{u}} > \lambda^{max}_{\mathbf{u}}$ or $\lambda_{\mathbf{v}} > \lambda^{max}_{\mathbf{v}}$, meaning that $\mathbf{v}^* = \mathbf{0}$ also. Thus, we either have that both $\mathbf{u}^*$ and $\mathbf{v}^*$ are zero or they are both are non-zero. By complementary slackness, (8) and (10), non-zero solutions imply that the inequality constraints of (4) are both tight and hold with equality; this proves part (iii). □

of Property 2. When $P_{\mathbf{u}}$ and $P_{\mathbf{v}}()$ are convex, notice that the SFPCA problem, (4), has (convex) quadratic inequality constraints and an objective that is bi-linear minus additive convex penalty functions. Thus, (4) is a concave problem in $\mathbf{u}$ with $\mathbf{v}$ fixed as well as the converse, meaning that it is *bi-concave*. □

of Theorem 2. Proof of part (i): Consider the penalized regression problem in (6). Some algebra shows that this can be arranged to give the following optimization problem:

$$\underset{\mathbf{u}}{\text{minimize}} \frac{1}{2} \mathbf{u}^T \mathbf{S}_{\mathbf{u}} \mathbf{u} - \mathbf{u}^T \mathbf{X} \mathbf{v} + \lambda_{\mathbf{u}} P_{\mathbf{u}}(\mathbf{u}). \tag{25}$$

Notice that the first two terms, let us call these $g(\mathbf{u})$, are convex and smooth (continuously differentiable) and the last term, $h(\mathbf{u})$, is non-smooth. To employ a proximal gradient scheme (Nesterov, 2005; Beck and Teboulle, 2009; Gong et al., 2013), we must show that the smooth part of this objective, $g(\mathbf{u})$, is gradient Lipschitz continuous. Consider that $\nabla g(\mathbf{u}) = \mathbf{S}_{\mathbf{u}} \mathbf{u} - \mathbf{X} \mathbf{v}$. Then, for some gradient Lipschitz constant, $L_{\mathbf{u}}$, the following must hold

$$\begin{aligned} ||\nabla g(\mathbf{u}) - \nabla g(\mathbf{u}')||_2 &\leq L_{\mathbf{u}} || \mathbf{u} - \mathbf{u}' ||_2 \\ || \mathbf{S}_{\mathbf{u}} \mathbf{u} - \mathbf{S}_{\mathbf{u}} \mathbf{u}' ||_2 &\leq L_{\mathbf{u}} || \mathbf{u} - \mathbf{u}' ||_2 \\ || \mathbf{S}_{\mathbf{u}} ||_2 || \mathbf{u} - \mathbf{u}' ||_2 &\leq L_{\mathbf{u}} || \mathbf{u} - \mathbf{u}' ||_2, \end{aligned}$$

and thus, the operator norm of $\mathbf{S}_{\mathbf{u}}$, $\lambda_{max}(\mathbf{S}_{\mathbf{u}})$ is a gradient Lipschitz constant. Then, the proximal mapping is defined as $\text{prox}_h(\mathbf{y}) = \text{argmin}_{\mathbf{x}} \{\frac{1}{2}||\mathbf{x} - \mathbf{y}||_2^2 + h(\mathbf{x})\}$. Note that this mapping exists for convex functions (Rockafellar, 1997) as well as non-convex functions that can be decomposed in to a difference of convex functions (Gong et al., 2013). This leads to the update step, $\mathbf{u}^{(t+1)} = \text{prox}_{\frac{\lambda_{\mathbf{u}}}{L_{\mathbf{u}}} h_{\mathbf{u}}}(\mathbf{u}^{(t)} - \frac{1}{L_{\mathbf{u}}} \nabla g(\mathbf{u}^{(t)}))$ (Beck and Teboulle, 2009), which is as given in Step 2 (a) of Algorithm 1. Monotonic convergence then follows from standard convergence analysis given in Nesterov (2005); Beck and Teboulle (2009) for convex penalty functions and in Gong et al. (2013) for non-convex penalties.

Proof of part (ii): Following the proof in part (i), the convergence rate of (convex) proximal gradient schemes with gradient Lipschitz constant, $L_{\mathbf{u}}$, follows from standard convergence analysis and developed and reviewed in Nesterov (2005); Beck and Teboulle (2009).

Proof of part (iii): The proof for $\mathbf{u}$ is given in Proposition 4 and the proof for $\mathbf{v}$ follows analogously.

Proof of part (iv): Notice that iteratively updating $\mathbf{u}$ and $\mathbf{v}$ as in Algorithm 1 is preforming block coordinate ascent. For these cyclical block updates, Tseng (2001) gives the conditions under which these schemes converge to a stationary point. In particular, the smooth part of the objective must follow certain regularity conditions. For our problem, the linear objective term of (4) as well as the quadratic constraint region are both continuously differentiable, and hence satisfy these regularity conditions. Next, the solution to each block update must be unique. Consider solving (4) with respect to $\mathbf{u}$, as given in (25). The gradient of the first two smooth terms is $\mathbf{S}_{\mathbf{u}} \mathbf{u} - \mathbf{X} \mathbf{v}$ yielding the Hessian $\mathbf{S}_{\mathbf{u}}$. As $\mathbf{S}_{\mathbf{u}} = \mathbf{I} + \alpha_{\mathbf{u}} \mathbf{\Omega}_{\mathbf{u}}$, it is positive definite, $\mathbf{S}_{\mathbf{u}} \succ 0$, by construction. The first terms are hence strictly convex, meaning that if $P_{\mathbf{u}}()$ is convex, (25) is strictly convex and any optimum is the unique global solution (Bertsekas et al., 2003). With $P_{\mathbf{v}}()$ convex, the same argument reveals that the problem in $\mathbf{v}$ is strictly convex, and hence the solution is also the unique. This then, satisfies the final condition from Tseng (2001). Therefore with convex penalties, the solution to Algorithm 1, $(\mathbf{u}, \mathbf{v})$, is a stationary point of (4).

It is worth noting the form of these stationary points. From Tseng (2001) if $(\mathbf{u}^*, \mathbf{v}^*)$ are the stationary points, then these are defined as:

$$\begin{aligned} \mathbf{X} \mathbf{v} - \lambda_{\mathbf{u}} \Gamma(\mathbf{u}) - 2\gamma_{\mathbf{u}} \mathbf{S}_{\mathbf{u}} \mathbf{u} \Big|_{\mathbf{u} = \mathbf{A}\, \mathbf{u}^*} &\leq 0 \ \forall \ \mathbf{A} \in \{\mathbf{A} : \mathbf{A}^T \mathbf{A} = \mathbf{I}\} \\ \mathbf{X}^T \mathbf{u} - \lambda_{\mathbf{v}} \Gamma(\mathbf{v}) - 2\gamma_{\mathbf{v}} \mathbf{S}_{\mathbf{v}} \mathbf{v} \Big|_{\mathbf{v} = \mathbf{B}\, \mathbf{v}^*} &\leq 0 \ \forall \ \mathbf{B} \in \{\mathbf{B} : \mathbf{B}^T \mathbf{B} = \mathbf{I}\}. \end{aligned}$$

As the left side of the above are the first order gradient conditions of (4), then this says that there are no directions in which $\mathbf{u}^*$ or $\mathbf{v}^*$ could move that would increase the objective. Note that this optimality result is much stronger that that of a mere local optima as shown for existing approaches to SPCA (Shen and Huang, 2008; Witten et al., 2009). □

of Property 3. From the proof of Theorem 2, we have shown that if both $P_{\mathbf{u}}()$ and $P_{\mathbf{v}}()$ are convex, then the sub-problems for $\mathbf{u}$ and $\mathbf{v}$ are strictly convex and the resulting solutions are unique. Thus, we can be assured that if we begin with a good initialization (i.e. the SVD), then our algorithm converges to a good stationary point that is unique. □

of Corollary 1. We must show that the convergence of the proximal gradient ascent scheme for both convex and non-convex penalties as shown in the proof of Theorem 2 also holds for what we term the positive proximal operator and non-negativity constraints. Consider optimizing with respect to $\mathbf{u} : \mathbf{u} \geq 0$ and $\mathbf{v} : \mathbf{v} \geq 0$. The non-negative orthant is a convex set and there trivially exists a strictly feasible point. Thus, from the arguments in the proof of Proposition 3, the KKT conditions are necessary for optimality of (5).

Next, we will show that Proposition 4 holds for the non-negative SFPCA problem and associated non-negative penalized regression problem. For optimizing (5) with respect to $\mathbf{u}$, the associated KKT conditions are

$$\begin{aligned} -2\gamma\,\mathbf{S}\,\mathbf{u} + \mathbf{X}\,\mathbf{v} - \lambda\Gamma(\mathbf{u}) - \delta = 0, \\ \gamma(\mathbf{u}^T\,\mathbf{S}\,\mathbf{u} - 1) = 0, \;\; -\delta\,\mathbf{u} = 0, \;\&\; \gamma, \delta \geq 0 \\ \mathbf{u}^T\,\mathbf{S}\,\mathbf{u} \leq 1, \;\&\; \mathbf{u} \geq 0. \end{aligned} \tag{26}$$

Here, δ is the dual variable associated with the non-negativity constraint; by complementary slackness, $\delta = 0$ if and only if $\mathbf{u} \geq 0$. Consider the non-negative penalized regression problem:

$$\underset{\mathbf{u}:\mathbf{u}\geq 0}{\text{minimize}} \;\; \frac{1}{2}||\,\mathbf{X}\,\mathbf{v} - \mathbf{u}\,||_2^2 + \lambda P(\mathbf{u}) + \frac{\alpha}{2}\,\mathbf{u}^T\,\mathbf{\Omega}\,\mathbf{u}\,. \tag{27}$$

The first-order KKT condition is

$$\mathbf{S}\,\mathbf{u} - \mathbf{X}\,\mathbf{v} + \lambda\Gamma(\mathbf{u}) + \delta = 0. \tag{28}$$

Notice that (28) is the same as (26) except for the scaling of 2γ on the first term. Thus, the arguments used in the proof of Proposition 4 can be used here, proving that the solution to (28) re-scaled to have $\mathbf{S}$-norm equal to one is the solution to (5) with respect to $\mathbf{u}$.

Now in addition, we must show that our so-called positive proximal operator and associated algorithm converges to the solution to (27). Using the same notation for the non-smooth part of our objective as in the proof of Theorem 2 part (i), replace the non-smooth part, $h(\mathbf{u}) = P(\mathbf{u})$ with $\tilde{h}(\mathbf{u}) = P(\mathbf{u}) + M\; I_{u\geq 0}$. Here, $I_{u\geq 0}$ is a non-negative indicator and M is some sufficiently large constant. As an indicator of a convex set is convex, then when $P()$ is convex, $\tilde{h}()$ is closed, convex, and separable; the convergence analysis for convex $P()$ follows from Beck and Teboulle (2009). For non-convex penalties with $P()$ satisfying assumption A3, $\tilde{h}()$ can easily be written as a difference of convex functions; thus, convergence analysis follows from Gong et al. (2013). Finally, we need to show that $\text{prox}_{\tilde{h}}(y)$ is equivalent to our positive proximal operator. For M sufficiently large, it is easy to see that $\text{prox}_{\tilde{h}}()$ enforces the non-negative constraint and is hence equivalent to the non-negatively constraint proximal operator, $\text{prox}_h^+(\mathbf{y}) = \text{argmin}_{\mathbf{x}:\mathbf{x}\geq 0}\{\frac{1}{2}||\mathbf{x} - \mathbf{y}||_2^2 + h(\mathbf{x})\}$. Thus, we have shown that Theorem 2 holds and establishes convergences results for the non-negatively constrained SFPCA problem. □

of Proposition 2. Consider optimizing (4) with respect to $\mathbf{u}$ when $P_{\mathbf{u}}(\mathbf{u}) = ||\,\mathbf{u}\,||_1$. In Proposition 4, we have shown that the solution for $\mathbf{u}$ is given by the solution to the ℓ_1 and generalized ℓ_2 penalized regression problem. This problem is then a generalization of the elastic net method of Zou and Hastie (2005). For this problem, Tibshirani and Taylor (2012) shows that the degrees of freedom can be estimated as the trace of the part of the projection matrix, $\mathbf{H} : \hat{Y} = \mathbf{H}Y$, associated with the non-zero or active variables. Consider calculating the projection matrix for the smooth part of our problem: $\frac{1}{2}||\,\mathbf{X}\,\mathbf{v} - \mathbf{u}\,||_2^2 + \frac{\alpha}{2}\,\mathbf{u}^T\,\mathbf{\Omega}\,\mathbf{u}$. The first-order condition is $-\mathbf{X}\,\mathbf{v} + \mathbf{u} + \alpha\,\mathbf{\Omega}\,\mathbf{u} = 0$, yielding solution, $\hat{\mathbf{u}} = (\mathbf{I} + \alpha\,\mathbf{\Omega})^{-1}\,\mathbf{X}\,\mathbf{v}$ and thus projection matrix $\mathbf{H} = (\mathbf{I} + \alpha\,\mathbf{\Omega})^{-1}$. Let $\mathcal{A}$ denote the indices of the non-zero elements of $\hat{\mathbf{u}}$, the solution to (6). Then, following from Tibshirani and Taylor (2012), we have that $\hat{df}(\hat{\mathbf{u}}) = \text{tr}[\mathbf{I}_{|\mathcal{A}|} - \alpha_{\mathbf{u}}\,\mathbf{\Omega}_{\mathbf{u}}(\mathcal{A}, \mathcal{A})]$ for ℓ_1 sparse penalties. The stated BIC criterion follows directly from its definition for the Gaussian likelihood associated with regression problems when the variance is assumed to be unknown. □

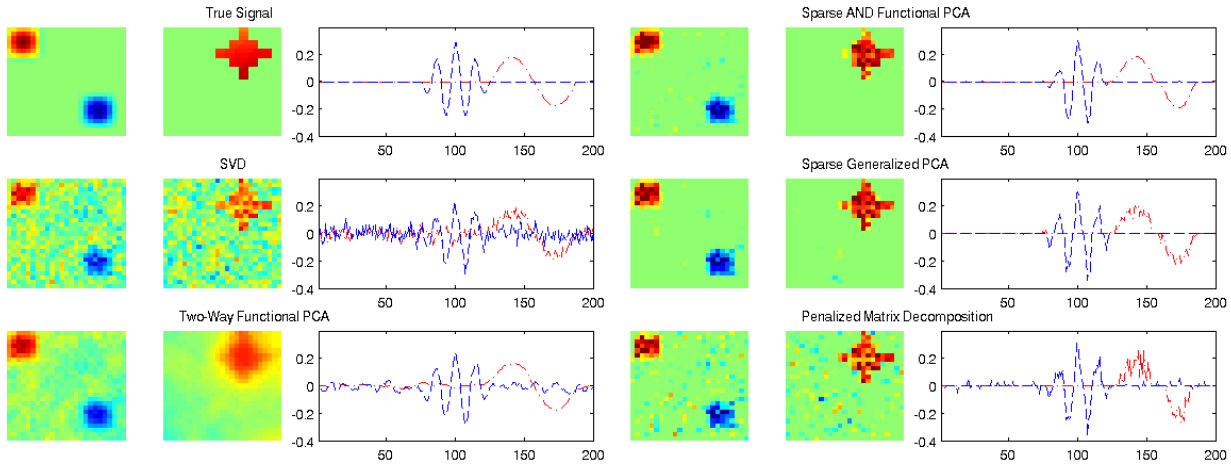


Figure 5: True spatial factors, $\mathbf{u}_1$ and $\mathbf{u}_2$ on a 25×25 grid, and temporal factors $\mathbf{v}_1$ (red, dotted-dashed), $\mathbf{v}_2$ (blue, dashed) in the top left panel, as well as those estimated by various methods for the two-way simulation study described in Appendix B.

		TWFPCA	SSVD	PMD	SGPCA ($\sigma = 1$)	SGPCA ($\sigma = 5$)	SFPCA
$\mathbf{u}_1$	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)
$\mathbf{v}_1$	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)
$\mathbf{u}_2$	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)
$\mathbf{v}_2$	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: Simulation results averaged over 50 replicates (with standard errors) in terms of 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). Best performing methods are boldfaced.

B Additional Simulation Results

We provide an additional simulation to study the performance of our SFPCA method for data with two-way factors. Data is simulated as described in Section 5.1 according to a rank 2 model. The true factors of this simulation are inspired by spatio-temporal neuroimaging data: $\mathbf{U} \in \Re^{625 \times 2}$ are the spatial factors on a 25×25 grid with $\mathbf{u}_1$ containing two spatially smooth regions of interest and $\mathbf{u}_1$ containing one region of interest with discrete edges; $\mathbf{V} \in \Re^{200 \times 2}$ are temporal factors as given by $\mathbf{v}_1$ and $\mathbf{v}_2$ in Section 5.1. These true factors are given in the top left panel of Figure 5. The signal-to-noise ratio is modulated by $d_1 = n/6$ and $d_2 = n/7$. We compare our SFPCA method to the same set of competing methods as described in Section 5.1. Additionally, the second differences matrix over a grid or a chain graph were used as $\mathbf{\Omega_u}$ and $\mathbf{\Omega_v}$ respectively; exponentially decaying smoothing matrices were used for the SGPCA method.

Simulation results for two-way signal recovery and feature selection of both the spatial and temporal factors are given in Table 2. The same evaluation metrics were used as described in Section 5.1. These results indicate that SFPCA performs best in recovering the first factor which is spatially smooth. The second factor, however, has sharp edges spatially and SFPCA does not perform as well at recovering these edges. Here, SGPCA performs slightly better, but SFPCA is still competitive, with better performance than other methods and with much improvement over the SVD. These results indicate that SFPCA performs well for recovering two-way sparse and smooth factors. As expected,

results are best for factors that are smooth with no discrete jumps or edges.

C Additional EEG Results

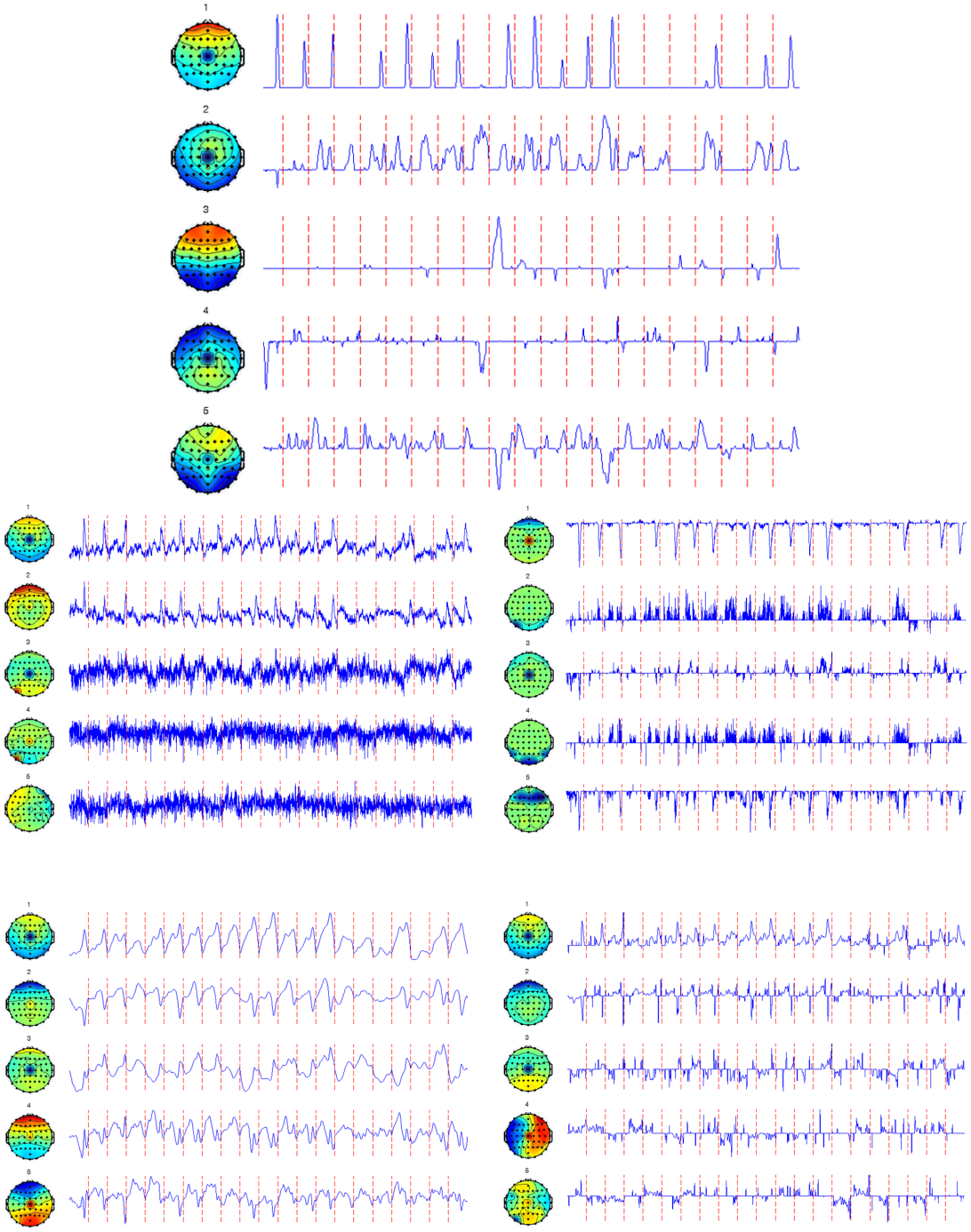


Figure 6: First five spatial and temporal SFPCA components (top),PCA components (middle left), two-way Sparse PCA components (middle right), two-way Functional PCA components (bottom left), and two-way Sparse Generalized PCA components (bottom right) for the EEG case study. While all methods identify some of the same temporal activity as seen in the peaks of the first two components, the resulting SFPCA components are much more interpretable and exhibit clearer temporal activation patterns.