SPARSE AND FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS

Genevera I. Allen

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

Michael Weylandt

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

ABSTRACT

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

Index Terms— regularized PCA, multivariate analysis

1. INTRODUCTION

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

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

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

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

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

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

2. A SPARSE AND FUNCTIONAL SINGULAR VALUE FORMULATION OF PCA

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

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

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

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

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

978-1-7281-0708-0/19/$31.00 ©2019 IEEE

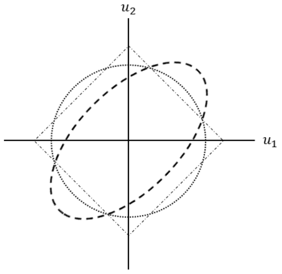

Fig. 1. Three constraints implicit in the ill-posed naive formulation of SFPCA: sparsity constraint (ℓ_1-ball), unit norm (ℓ_2-ball), and smoothness (elliptical region). In general, it is difficult for a point to lie on the boundary of all three regions simultaneously, leading to degenerate solutions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

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

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

Surprisingly, this natural generalization fails, often spectacularly!

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. COMPUTATION OF SPARSE AND FUNCTIONAL PRINCIPAL COMPONENTS

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

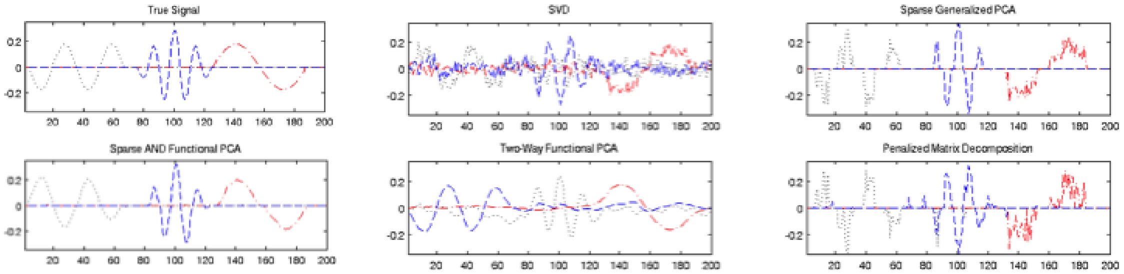

Fig. 2. Simulated factors used for the simulation study in Section 4 and estimates thereof: $\boldsymbol{v}_1$ (red, dotted-dashed); $\boldsymbol{v}_2$ (blue, dashed); and $\boldsymbol{v}_3$ (black, dotted). Only SFPCA is able to simultaneously identify the spatial sparsity and smooth structure of the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

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

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

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

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

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

$$\operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{B}^n_{\boldsymbol{S_u}}} \frac{1}{2}\|\boldsymbol{X v} - \boldsymbol{u}\|_2^2 + \lambda_{\boldsymbol{u}} P_{\boldsymbol{u}}(\boldsymbol{u}) + \frac{\alpha_{\boldsymbol{u}}}{2}\boldsymbol{u}^T\boldsymbol{\Omega_u u}. \quad (4)$$

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

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

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

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

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

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

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

3.1. Selection of Regularization Parameters

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

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

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

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

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

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

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

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

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

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

4. SIMULATION STUDY

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

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

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

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

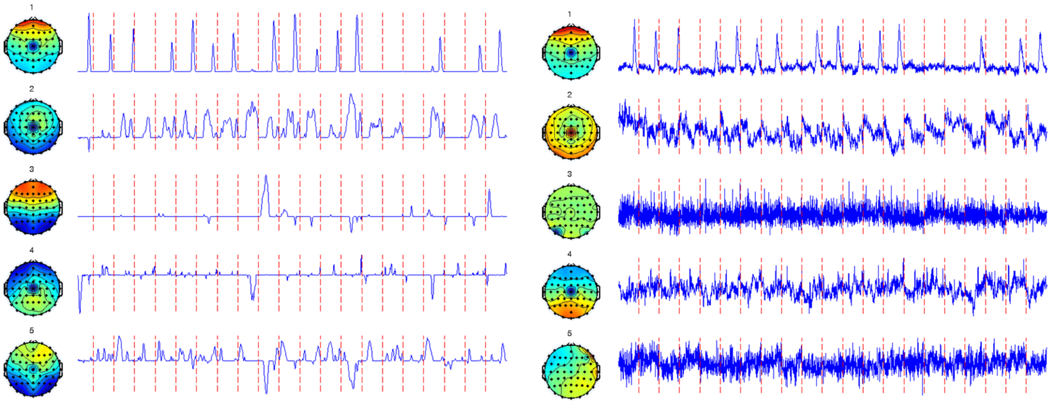

Fig. 3. EEG Case Study: first five spatial and temporal SFPCA components (left) and ICA components (right). While SFPCA and ICA identify similar structures in the first two components, the temporal sparsity of the SFPCA components makes them more readily interpretable. Additionally, the SFPCA finds structure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

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

5. CASE STUDY: EEG DATA

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

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

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

6. DISCUSSION

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

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

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

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

7. REFERENCES

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

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

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

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

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

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

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

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

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

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

[11] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 67, no. 1, pp. 91–108, 2005. DOI: 10.1111/j.1467-9868.2005.00490.x.

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

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

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

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

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

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

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

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

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

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

[22] D. Dua and E. Karra Taniskidou, *UCI machine learning repository*. Available at: http://archive.ics.uci.edu/ml.

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

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

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

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