*The Annals of Statistics*
2015, Vol. 43, No. 1, 299–322
DOI: 10.1214/14-AOS1273


# SPARSISTENCY AND AGNOSTIC INFERENCE IN SPARSE PCA

BY JING LEI[1] AND VINCENT Q. VU

*Carnegie Mellon University and The Ohio State University*

The presence of a sparse "truth" has been a constant assumption in the theoretical analysis of sparse PCA and is often implicit in its methodological development. This naturally raises questions about the properties of sparse PCA methods and how they depend on the assumption of sparsity. Under what conditions can the relevant variables be selected consistently if the truth is assumed to be sparse? What can be said about the results of sparse PCA without assuming a sparse and unique truth? We answer these questions by investigating the properties of the recently proposed Fantope projection and selection (FPS) method in the high-dimensional setting. Our results provide general sufficient conditions for sparsistency of the FPS estimator. These conditions are weak and can hold in situations where other estimators are known to fail. On the other hand, without assuming sparsity or identifiability, we show that FPS provides a sparse, linear dimension-reducing transformation that is close to the best possible in terms of maximizing the predictive covariance.

**1. Introduction.** Sparse principal components analysis (PCA) is a relatively new and popular technique for simultaneous dimension reduction and variable selection in high-dimensional data analysis [e.g., Jolliffe, Trendafilov and Uddin (2003), Zou, Hastie and Tibshirani (2006)]. It combines the central idea of classic (or ordinary) PCA [Hotelling (1933), Pearson (1901)] with the notion of sparsity: it seeks linear transformations that reduce the dimension of the data, while depending on a small number of variables, but retain as much variation as possible. In the population setting, these linear transformations correspond to the projectors of the $k$-dimensional principal subspaces, spanned by the eigenvectors of the population covariance matrix. The appeal of sparsity is that it not only enhances interpretability, but it can yield consistent estimates when sparsity is truly present in the population, even in high dimensions [Johnstone and Lu (2009)].

The development of sparse PCA has taken a brisk pace over the past decade. Methodological developments include regularized estimators based on penalizing or constraining the variance maximization formulation of PCA [Jolliffe, Trendafilov and Uddin (2003), Journée et al. (2010), Witten, Tibshirani and Hastie (2009)], regression or low-rank approximation [Shen and Huang (2008),

Received January 2014; revised September 2014.

[1]Supported by NSF Grant BCS-0941518, NSF Grant DMS-14-07771 and NIH Grant MH057881.

*MSC2010 subject classifications.* 62H12.

*Key words and phrases.* Principal components analysis, subspace estimation, sparsity, variable selection, agnostic inference.

Zou, Hastie and Tibshirani (2006)], convex relaxations [d'Aspremont, Bach and El Ghaoui (2008), d'Aspremont et al. (2007), Vu et al. (2013)], two-stage procedures based on diagonal thresholding [Johnstone and Lu (2009), Paul and Johnstone (2012)] and algorithmic variations of iterative thresholding [Ma (2013), Yuan and Zhang (2013)]. Theoretical developments including consistency, rates of convergence, minimax risk bounds for estimating eigenvectors and principal subspaces and detection have been established under various statistical models [Amini and Wainwright (2009), Berthet and Rigollet (2013a), Cai, Ma and Wu (2013), Johnstone and Lu (2009), Lounici (2013), Ma (2013), Vu and Lei (2012, 2013), Vu et al. (2013)].

The presence of a sparse "truth" has been an explicit assumption in the theoretical analysis of sparse PCA and is often an implicit assumption in its methodological development. Here the "truth" refers to the leading $k$-dimensional principal subspace. This naturally raises questions about the properties of sparse PCA methods and how they depend on the assumption of sparsity. Under what conditions can the relevant variables be selected consistently if the truth is assumed to be sparse? If the truth is not sparse, and/or not unique, what can be said about the results of sparse PCA? The first question is essentially concerned with variable selection consistency, or *sparsistency*. The second question is a bit more slippery, because it essentially requires us to assume nothing beyond independence of the observations. In other words, the second question is concerned with *agnostic inference properties* of an estimation method. In this paper, we investigate variable selection consistency and agnostic inference properties of the recently proposed *Fantope projection and selection* (FPS) method due to Vu et al. (2013).

FPS formulates the sparse PCA problem as a semidefinite program (SDP) whose solution is a sparse estimate of the projector of the principal subspace. It extends the so-called DSPCA formulation of d'Aspremont et al. (2007) from the one-dimensional ($k = 1$) case to the multidimensional ($k > 1$) case, and it presents a change in perspective by focusing on projectors rather than individual eigenvectors. FPS is appealing for both theoretical and computational reasons. Since it directly estimates the projector of the $k$-dimensional principal subspace, there is no need for iterative deflation [e.g., Mackey (2009)], and hence an SDP need only be solved once rather than $k$ separate times as in DSPCA. Vu et al. (2013) developed an efficient alternating direction method of multipliers algorithm to compute FPS, and established $\ell_2$ consistency of FPS under very mild conditions on the population and input matrices. Most notably, FPS does not require the stringent spiked covariance model assumption (i.e., the population covariance matrix is a sparse low-rank matrix plus identity) that is required by many competing methods such as diagonal thresholding. This makes FPS applicable to a much wider range of problems, including the important case of correlation matrices where diagonal thresholding cannot even be used. (See Section 3 for another example.) However, the variable selection and agnostic inference properties of FPS remain unknown.

Sparsistency is the ability of an estimator to accurately select the correct subset of variables when applied to a random sample generated from a model where only a subset of variables is assumed to be relevant. Conditions under which sparsistency holds provide important insights about both the estimator and the model. They have been studied extensively in other high-dimensional inference problems such as linear regression [Fan and Li (2001), Meinshausen and Bühlmann (2006), Wainwright (2009), Zhao and Yu (2006)] and Gaussian graphical model selection [Lam and Fan (2009), Ravikumar et al. (2011), Rothman et al. (2008)]. In contrast, theoretical analyses of sparse PCA have mainly focused on consistency and rates of convergence in matrix norm, with relatively less progress on variable selection. An exception is Amini and Wainwright (2009), who analyzed DSCPCA under a stringent spiked covariance model with $k = 1$, where the population covariance matrix is block diagonal and its leading eigenvector is assumed to have a small number of nonzero entries of constant magnitude. Their work is an important first step, but it leaves open whether or not their stringent conditions can be loosened and it also does not address the $k > 1$ case.

In the first part of this paper, we investigate the sparsistency of FPS under general conditions. Our main results (Theorems 1 and 2) give broad sufficient conditions under which FPS can exactly recover the relevant variables. Roughly, the conditions are that (1) the relevant variables are not too correlated with the irrelevant variables (limited correlation), and (2) the leverages (diagonals of the projector) of the relevant variables are large enough. Interestingly, these conditions are analogous to so-called (1) "irrepresentability" and (2) "$\beta$-min" conditions for variable selection consistency of the Lasso [Bühlmann and van de Geer (2011), Meinshausen and Bühlmann (2006), Zhao and Yu (2006)]. To our knowledge, this is the first sparsistency result for principal subspaces. When $k = 1$, it generalizes the results of Amini and Wainwright (2009) in several directions, the most important of which is that it relaxes their block-diagonal condition on the population covariance matrix.

The second part of this paper addresses the question of assumption-free interpretation of sparse PCA within a framework that we call *agnostic inference*. Our goal is to provide both analysis and interpretation of sparse PCA with essentially no assumptions beyond independence of observations. The terminology is borrowed from the learning theory literature where the chief concern is estimating a classifier or regression function without assumptions on the model [Kearns, Schapire and Sellie (1994)]; however, much of our perspective is influenced by earlier work on maximum likelihood under misspecification [Berk (1966), Huber (1967), White (1982)], interpretations obtained by extending the maximum likelihood principle [Akaike (1973)], and the notion of *persistence* of high-dimensional linear predictors proposed by Greenshtein and Ritov (2004). Our point is that although FPS is derived under the assumption of sparsity, its results can still be interpreted even when sparsity does not hold. The main result (Theorem 3) is that without assuming sparsity or identifiability, FPS provides a sparse, linear dimension-reducing transformation that is close to the best possible in terms of maximizing the *predictive covariance*.

The remainder of the paper is organized as follows. Section 2 provides the technical background and conditions that are necessary to state our results—divided between Sections 3 (sparsistency) and 4 (agnostic inference). We discuss these results in Section 5, and defer their proofs to the Appendix. Finally, we collect our notation below for our readers' convenience.

*Notation.* For two matrices $A, B$ with conformable dimensions, $\langle A, B\rangle := \operatorname{trace}(A^T B)$ denotes the trace inner product. For a vector $v \in \mathbb{R}^k$ and $q \in [0, \infty]$, $\|v\|_q = (\sum_{i=1}^p |v_i|^q)^{1/q}$ is the $\ell_q$ norm if $0 < q < \infty$; when $q = 0$, $\|v\|_0$ is the number of nonzero entries of $v$; when $q = \infty$, $\|v\|_\infty = \max_{1\le i\le k} |v_i|$. For a matrix $A \in \mathbb{R}^{n\times m}$, and index sets $J_1 \subseteq [n]$ and $J_2 \subseteq [m]$, $A_{J_1 J_2}$ denotes the $|J_1| \times |J_2|$ submatrix of $A$ consisting of rows in $J_1$ and columns in $J_2$, and $A_{J_1 *}$ ($A_{*J_2}$) denotes the submatrix consists of corresponding rows (columns). Given $q_1, q_2 \in [0, \infty]$ and $A \in \mathbb{R}^{n\times m}$, the matrix $(q_1, q_2)$-pseudonorm $\|A\|_{q_1,q_2}$ is defined as $(\|A_{1*}\|_{q_1}, \|A_{2*}\|_{q_1}, \dots, \|A_{n*}\|_{q_1})_{q_2}$. As usual, the spectral norm of $A$ is denoted $\|A\|$ and the Frobenius norm is $\|A\|_F := \langle A, A\rangle^{1/2}$. If $A$ is a symmetric matrix, $\lambda_j(A)$ denotes the $j$th largest eigenvalue of $A$. We will use $\Sigma$ to denote the $p \times p$ underlying true covariance matrix, whose ordered eigenvalues are $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$. For a square matrix $A$, $\operatorname{diag}(A)$ denotes its diagonal vector. For a vector $v$, $\operatorname{supp}(v)$ is the support of $v$ (the index set corresponding to nonzero entries).

## 2. Preliminaries.

Let $\Sigma \in \mathbb{R}^{p\times p}$ be a symmetric matrix with spectral decomposition

$$\Sigma = \sum_{j=1}^{p} \lambda_j u_j u_j^T,$$

where $\lambda_1 \ge \cdots \ge \lambda_p$ are eigenvalues and $u_1, \dots, u_p \in \mathbb{R}^p$ is an orthonormal basis of eigenvectors. The $k$-dimensional principal subspace of $\Sigma$ is the subspace spanned by $u_1, \dots, u_k$. It is unique if and only if the spectral gap $\lambda_k - \lambda_{k+1} > 0$, and its projector (orthogonal projection matrix) is

$$\Pi = \sum_{j=1}^{k} u_j u_j^T = UU^T,$$

where $U$ is the orthonormal matrix with columns $u_1, \dots, u_k$. Every subspace has a unique projector and so we will consider the principal subspace and $\Pi$ to be equivalent, and we will also assume that $k$ is known or fixed in advance.

2.1. *Sparse principal subspaces.* Estimation of the principal subspace requires at minimum that it be well-defined. When this is the case, we can consider

$\Pi$ to be a mapping $x \mapsto \Pi x$ and so it makes sense to consider indices of the variables that $\Pi$ depends on. Since $\Pi$ is positive semidefinite, this is equivalent to the indices of the nonzero diagonal entries of $\Pi$, because row/column $i$ of $\Pi$ is nonzero if and only if $\Pi_{ii} \neq 0$.

CONDITION 1 (SPS). $\Sigma$ satisfies the *sparse principal subspace* condition with support set $J$ if

$$\text{(SPS)} \qquad \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma) > 0 \quad \text{and} \quad \operatorname{supp}\bigl(\operatorname{diag}(\Pi)\bigr) = J.$$

(SPS) is the minimal requirement for sparse principal subspace estimation, and the assumption will only be used in Section 3 in our investigation of sparsistency. The spectral gap condition ensures that the principal subspace is identifiable, and the support set definition states that the principal subspace does not depend on variables outside of $J$. This corresponds to a notion of subspace sparsity introduced by Vu and Lei (2013) called $\ell_0$ *row sparsity*, and it can be shown that $J = \bigcup_{j=1}^{k} \operatorname{supp}(u_j)$ for any orthonormal basis $\{u_1, \dots, u_k\}$ of the principal subspace [Vu and Lei (2013)].

2.2. *Input matrix accuracy.* When (SPS) is assumed, the main statistical inference problem considered in this paper is, in a general setting, to estimate $J$ from a symmetric noisy version $S$ of $\Sigma$. We then extend the interpretation and analytical properties of sparse PCA solutions without assuming (SPS). In both parts, the estimation accuracy depends on the noisiness of $S$ as an approximation to $\Sigma$, which will be quantified by an entrywise tail bound on

$$W := S - \Sigma.$$

As motivated by principal component analysis, it may be helpful to think of $\Sigma$ as the covariance of a $p$-dimensional random vector and $S = S_n$ as sample covariance matrix of a random sample of size $n$, but that is not strictly necessary for our theoretical analysis. In fact, our sparsistency results do not even have to assume that $\Sigma$ or $S$ are positive semidefinite. In the following, we describe two probabilistic models that imply a strong entrywise tail bound on $W$.

EXAMPLE 1 (Sample covariance). Let $X, X_1, X_2, \dots, X_n \in \mathbb{R}^p$ be i.i.d. random vectors with $\operatorname{Var}(X) = \Sigma \succeq 0$ and let $S$ be the sample covariance matrix:

$$S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T,$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. We assume throughout this paper that

$$\log p \leq n.$$

By Bernstein's inequality [van der Vaart and Wellner (1996), Chapter 2.2] if $X$ has sub-Gaussian tails in that there exists constants $K, C > 0$ such that

$$\mathbb{P}(|v^T(X - \mathbb{E}X)| \geq t) \leq K \exp[-Ct^2/(v^T \Sigma v)] \qquad \text{for all } v \neq 0, \tag{1}$$

then there is an absolute constant $c > 0$ such that $S$ satisfies, for $\sigma \geq c\lambda_1$,

$$\mathbb{P}\left(\|W\|_{\infty,\infty} \geq \sigma\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-2}. \tag{2}$$

In other words, the maximum entrywise error is bounded by $\sigma\sqrt{\log p/n}$ with high probability. This fact will be the starting point of subsequent analysis of the sparsistency of the FPS estimator introduced in Section 2.3. The tail bound (2) is well known and a proof of a stronger result that implies (2) can be found in Vu and Lei (2012), Lemma 3.2.2.

EXAMPLE 2 (Random graph models). Here, we give an example that does not involve an i.i.d. random sample and the rate of error bound on $S$ depends only on $p$. Consider a random graph model with $p$ nodes where edges appear independently with probability $c_{ij}$ for all $1 \leq i < j \leq p$. Let $A$ be the random adjacency matrix such that $A_{ij} = \pm 1$ according to the presence/absence of edge, then the pair $S = AA^T/(p-1)$ and $\Sigma = \mathbb{E}S$ satisfies

$$\mathbb{P}\left(\|S - \Sigma\|_{\infty,\infty} \geq c\sqrt{\frac{\log p}{p-1}}\right) \leq 2p^{-2}$$

for some universal constant $c$.

This model is related to the planted clique problem where $c_{ij} = 1$ for all $1 \leq i < j \leq s$, and $c_{ij} = 1/2$ everywhere else. The leading eigenvector of $\Sigma$ is $(1/\sqrt{s}, \ldots, 1/\sqrt{s}, 0, \ldots, 0)$ and it is supported on $J = \{1, \ldots, s\}$. Our main result implies that FPS finds the planted clique with high probability when $s \geq c\sqrt{p \log p}$ for some absolute constant $c$. This is within a factor of $\sqrt{\log p}$ of the best known result for polynomial time recovery in the planted clique problem [Deshpande and Montanari (2013)]. Berthet and Rigollet (2013b) give another reduction of the planted clique model to a sparse PCA problem.

For simplicity of notation and presentation, we will focus on the case of sample covariance matrix in the rest of this paper. But most of our sparsistency results are applicable to a broader range of problems as exemplified in the random graph example.

2.3. *Fantope projection and selection.* Vu et al. (2013) recently proposed an estimator for $\Pi$, called *Fantope projection and selection* (FPS), defined as a solution $\widehat{H}$ to the following semidefinite program:

$$\widehat{H} := \arg\max\{\langle S, H\rangle - \rho\|H\|_{1,1}\} \quad \text{subject to} \quad H \in \mathcal{F}^k, \tag{3}$$

where

$$\mathcal{F}^k := \{H : 0 \preceq H \preceq I \text{ and } \operatorname{trace}(H) = k\}$$

is the trace-$k$ *Fantope*, $k > 0$, and $\rho \geq 0$ is a tuning parameter. Vu et al. (2013) showed that FPS can be efficiently computed by alternating direction method of multipliers [ADMM, e.g., Boyd et al. (2010)]. When $\rho = 0$, a solution is given by the projector of the $k$-dimensional principal subspace of $S$ (see Lemma 1 below). The $\ell_1$ penalty term encourages the solution to be sparse. Moreover, the decomposability of the $\ell_1$ penalty term [Negahban et al. (2012)] makes it straightforward to analyze the statistical properties of FPS. In particular, Vu et al. (2013) established a near-optimal Frobenius norm error bound for the FPS estimator under general conditions. In the next section, we will show that, if $\Sigma$ satisfies the (SPS) and $S$ satisfies the maximum error bound assumption (2), then under mild conditions, $\operatorname{supp}[\operatorname{diag}(\widehat{H})] = J$ with high probability for appropriate choices of $\rho$.

In general, the solution to (3), and hence the FPS estimator may not be unique. However, we will show that it is unique with high probability when the (SPS) and maximum error bound assumption hold. The argument utilizes the following elastic net version of FPS:

$$\max\left\{\langle S, H\rangle - \rho\|H\|_{1,1} - \frac{\tau}{2}\|H\|_F^2\right\} \quad \text{subject to} \quad H \in \mathcal{F}^k. \tag{4}$$

Since the objective is a strongly concave function, the solution of (4) is unique. A very interesting and important fact is that when $\rho$ and $\tau$ are small enough, if a solution of (3) is sparse then it must be the unique solution of (4). This observation will be proved in the Appendix and play a key role in establishing the uniqueness of solution for the original FPS problem.

We conclude this section by introducing some basic properties of the Fantope, which will be used repeatedly in the proof of main results. Further properties and discussion of the Fantope will be given in Section 4. Denote the Euclidean projection of a $p \times p$ symmetric matrix $A$ onto $\mathcal{F}^k$ by

$$\mathcal{P}_{\mathcal{F}^k}(A) := \operatorname*{arg\,min}_{Z \in \mathcal{F}^k} \|A - Z\|_F^2.$$

LEMMA 1 (Basic properties of Fantope projection). *Let $A$ be a symmetric matrix with eigenvalues $\gamma_1 \geq \cdots \geq \gamma_p$ and orthonormal eigenvectors $v_1, \ldots, v_p$.*

1. $\max_{H \in \mathcal{F}^k} \langle A, H\rangle = \gamma_1 + \cdots + \gamma_k$ *and the maximum is achieved by the projector of a $k$-dimensional principal subspace of $A$. Moreover, the maximizer is unique if and only if $\gamma_k > \gamma_{k+1}$.*
2. $\mathcal{P}_{\mathcal{F}^k}(A) = \sum_j \gamma_j^+(\theta) v_j v_j^T$, *where* $\gamma_j^+(\theta) = \min(\max(\gamma_j - \theta, 0), 1)$ *and $\theta$ satisfies the equation $\sum_j \gamma_j^+(\theta) = k$.*

3. *If* $0 < \tau \le \gamma_k - \gamma_{k+1}$, *then*

$$\arg\max_{H\in\mathcal{F}^k}\langle A, H\rangle = \arg\max_{H\in\mathcal{F}^k}\langle A, H\rangle - \frac{\tau}{2}\|H\|_F^2 = \mathcal{P}_{\mathcal{F}^k}(\tau^{-1}A) = \sum_{j=1}^k v_j v_j^T,$$

*uniquely.*

A proof of Lemma 1 is given in Section A.2.

**3. Sparsistency.** Throughout this section, we assume that $\Sigma$ satisfies (SPS) with dimension $k$ and support set $J = \{1, 2, \dots, s\}$ for some $s \ll p$, and that $S$ satisfies the maximum error bound condition (2) with some $\sigma > 0$. The sample covariance matrix is covered as a special case in view of Example 1.

Intuitively, variable selection would be easier if the relevant variables (those in $J$) and noise variables (those in $J^c$) are not too correlated. In the context of sparse linear regression, such an intuition leads to the famous Irrepresentable Condition [Wainwright (2009), Zhao and Yu (2006)]. In sparse subspace estimation, we have the analogous Limited Correlation Condition (LCC). In order to state the condition concisely, we use the following block representation of $\Sigma$:

$$\Sigma = \begin{pmatrix} \Sigma_{JJ} & \Sigma_{JJ^c} \\ \Sigma_{J^cJ} & \Sigma_{J^cJ^c} \end{pmatrix}.$$

Similar block representations can be defined for $S$ and $W = S - \Sigma$.

Our main technical condition, the limited correlation condition (LCC) is given below.

CONDITION 2 (LCC). A symmetric matrix $\Sigma$ satisfies the *limited correlation condition* with constant $\alpha \in (0, 1]$ if

$$\text{(LCC)} \qquad \frac{8s}{\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)}\|\Sigma_{J^cJ}\|_{2,\infty} \le 1 - \alpha.$$

(LCC) contains the condition assumed by Amini and Wainwright (2009) as a special case, where $\Sigma_{J^cJ} = 0$, and hence (LCC) holds with $\alpha = 1$. Another popular model for sparse PCA is the *spiked covariance model*, where $\lambda_k(\Sigma_{JJ}) \ge c$, $\Sigma_{J^cJ^c} = cI_{p-s}$, and $\Sigma_{J^cJ} = 0$. An important difference between (LCC) and the assumptions in previous works is that previous assumptions, for example, the spiked covariance model, usually imply that the relevant variables can be selected with good accuracy by thresholding the diagonal entries, while (LCC) contains situations where such diagonal thresholding intuition does not work. Here, we illustrate this difference by a toy example with $p = 3$, $k = 1$, $J = \{1, 2\}$:

$$\Sigma = \begin{pmatrix} 0.9 & 0.8 & t \\ 0.8 & 0.9 & -t \\ t & -t & 1 \end{pmatrix}. \tag{5}$$

This $\Sigma$ satisfies (LCC) with $\alpha = 0.3$ for any $|t| \leq 0.02$, but picking large diagonal entries of $\Sigma$ does not select the relevant variables.

To our knowledge, the (LCC) is the first sufficient condition for consistent sparse PCA variable selection without assuming $\Sigma$ being block-diagonal and is also the first sufficient condition for sparse subspace variable selection consistency.

3.1. *Sparsistency of FPS.* We state two versions of our main results. The first is a more general, deterministic result that provides sufficient conditions for uniqueness, false positive control, and false negative control of $\mathrm{supp}(\widehat{H})$. The second specializes the general result to the case where $S$ satisfies an entrywise error bound (2) like the sample covariance matrix in Example 1, and provides probabilistic guarantees for sparsistency of FPS.

THEOREM 1 (Deterministic support recovery). *Assume* $\Sigma$ *satisfies* (SPS). *If the FPS penalty parameter* $\rho$ *satisfies*

$$\rho^{-1}\|S - \Sigma\|_{\infty,\infty} + \frac{8s}{\lambda_k - \lambda_{k+1}}\|\Sigma_{J^cJ}\|_{2,\infty} \leq 1 \tag{6}$$

*and*

$$0 < \lambda_k - \lambda_{k+1} - 4\rho s\left(1 + \frac{8\lambda_1}{\lambda_k - \lambda_{k+1}}\right), \tag{7}$$

*then the solution* $\widehat{H}$ *of FPS problem* (3) *is unique and satisfies* $\mathrm{supp}(\mathrm{diag}(\widehat{H})) \subseteq J$. *If in addition, either*

$$\min_{j \in J} \sqrt{\Pi_{jj}} > \frac{4\rho s}{\lambda_k - \lambda_{k+1}} \quad \textit{or} \tag{8}$$

$$\min_{(i,j)\in J^2} |\Sigma_{ij}| > 2\rho \quad \textit{and} \quad \mathrm{rank}\big(\mathrm{sign}(\Sigma_{JJ})\big) = 1, \tag{9}$$

*then the FPS solution satisfies* $\mathrm{supp}(\mathrm{diag}(\widehat{H})) = J$.

Theorem 1 consists of two parts. The first part provides a set of sufficient conditions [(6) and (7)] for no false positives. The second part gives two additional conditions that individually guarantee no false negatives, and hence exact recovery. We discuss these parts separately.

*False positive control.* (6) reveals the motivation for (LCC). When (LCC) holds, one can choose $\rho = \|S - \Sigma\|_{\infty,\infty}/\alpha$ so that (6) holds. On the other hand, (7) puts some upper bound constraint on $\rho$. When $S$ is random and satisfies the maximum error bound condition (2), $\|S - \Sigma\|_{\infty,\infty}$ depends on $(n, p, \sigma)$. Then (6) and (7) jointly put a constraint on $(s, p, n, \sigma, \lambda_1, \lambda_k, \lambda_{k+1})$ so that there exists a $\rho$ satisfying both conditions.

(7) puts an upper bound on the sparsity penalty parameter $\rho$. It may seem counterintuitive since a larger value of $\rho$ will lead to a sparser solution. In fact, $\rho$ cannot be too large because otherwise the $\ell_1$ penalty term will outweigh the PCA objective in the FPS problem, leading to a large estimation bias. Consider the example given in (5) with $t = 0$, if $S = \Sigma$ and $\rho > 0.9$; the FPS solution will return a projection matrix corresponding to eigenvector $(0, 0, 1)$, which is supported outside of the true subset. In general, when $\rho \to \infty$, the FPS solution will be a diagonal matrix taking value 1 on diagonal entries corresponding to the $k$ largest diagonal entries of $S$, and 0 elsewhere.

The proof of false positive control in Theorem 1, as given in Section A.1, consists of two main steps. The first step (Section A.1.1) is to show that there exists a solution of the FPS problem (3) supported on $J$, using the primal–dual witness (PDW) argument [Amini and Wainwright (2009), Ravikumar et al. (2011), Wainwright (2009)]. The PDW argument first constructs a sparse solution $\tilde{H}$ supported on $J$ by solving the FPS problem (3) under additional sparsity constraint $\operatorname{supp}[\operatorname{diag}(H)] \subseteq J$. Then it is shown that when $\rho$ is large enough, with high probability one can find a dual variable $\widehat{Z}$ such that the primal–dual pair $(\tilde{H}, \widehat{Z})$ satisfies the KKT condition, and hence is optimal for the original problem. When the solution is unique, this ensures that the optimizer is supported on $J$. The challenge here is to establish KKT condition when $\Sigma$ is not block diagonal, which requires a careful and delicate subspace perturbation analysis in comparing the FPS solution and the population projector (Lemmas 2 and 3).

The second step is to show that, under the conditions assumed in the theorem, the sparse solution constructed in the primal–dual witness argument is indeed rank-$k$ and also unique. Our proof of uniqueness is novel and makes use of the elastic net version of FPS (4). A key fact used in the proof is that, for small enough values of $\tau$, the two problems have the same solution and the uniqueness of FPS solution follows essentially from that of the elastic net version. The details are given in Section A.1.2.

*False negative control.* Having established false positive control in Theorem 1, full sparsistency will be established if we can show that the number of false negatives is also zero. In sparsity pattern recovery, the number of false negatives is typically controlled by assuming a lower bound on the magnitude of signals carried by relevant variables. In the context of principal subspace estimation, our first sufficient condition for false negative control (8) originates from a Frobenius norm error bound of FPS established in Vu et al. (2013):

$$\|\widehat{H} - \Pi\|_F \le \frac{4\rho s}{\lambda_k - \lambda_{k+1}}. \tag{10}$$

The other sufficient condition for controlling false negative (9) is motivated by an assumption used by Amini and Wainwright (2009) for the $k = 1$ case where the leading eigenvector is assumed to be $v_1 = (\mathbf{1}_s/\sqrt{s}, 0)$ (where $\mathbf{1}_s$ is the $s \times 1$

vector of ones, and the signs of nonzero entries can actually be arbitrary) and $\Sigma_{JJ} = \theta v_1 v_1^T + I_s$. Let $\operatorname{sign}(\Sigma_{JJ})$ be the $s \times s$ matrix of entry-wise signs of $\Sigma_{JJ}$. Our condition (9) generalizes that of Amini and Wainwright (2009) in three directions. First, we allow principal subspaces of dimension $k > 1$. Second, we allow nonzero correlation between the relevant and irrelevant variables, whereas Amini and Wainwright (2009) assumes a block diagonal structure. Third, we do not require a generalized spiked covariance model as in Amini and Wainwright (2009). The proof of the second part of Theorem 1 is given in Section A.2.

THEOREM 2 (Sparsistency). *Assume that $\Sigma$ satisfies* (SPS) *and* (LCC), *and that $S$ satisfies the maximum error bound* (2) *with scaling factor $\sigma$. If*

$$s\sqrt{\frac{\log p}{n}} < \frac{\alpha(\lambda_k - \lambda_{k+1})^2}{4\sigma(8\lambda_1 + \lambda_k - \lambda_{k+1})}, \tag{11}$$

*and the FPS penalty parameter $\rho$ in* (3) *satisfies*

$$\rho = \frac{\sigma}{\alpha}\sqrt{\frac{\log p}{n}},$$

*then with probability at least $1 - 2p^{-2}$, the FPS estimate $\widehat{H}$ is unique and satisfies* $\operatorname{supp}(\operatorname{diag}(\widehat{H})) \subseteq J$. *If in addition, either*

$$\min_{j \in J} \sqrt{\Pi_{jj}} > \frac{4s\sigma}{\alpha(\lambda_k - \lambda_{k+1})}\sqrt{\frac{\log p}{n}} \quad \textit{or} \tag{12}$$

$$\min_{(i,j) \in J^2} \Sigma_{ij} > \frac{2\sigma}{\alpha}\sqrt{\frac{\log p}{n}} \quad \textit{and} \quad \operatorname{rank}\big(\operatorname{sign}(\Sigma_{JJ})\big) = 1, \tag{13}$$

*then* $\operatorname{supp}(\operatorname{diag}(\widehat{H})) = J$.

PROOF. Using the maximum error bound condition, with probability at least $1 - 2p^{-2}$ we have $\rho^{-1}\|S - \Sigma\|_{\infty,\infty} \le \alpha$. This together with the property (LCC) of $\Sigma$ establishes (6). On the other hand, (11) ensures that (7) holds. On the other hand, (12) implies (8), and (13) implies that the choice of $\rho$ satisfies (9). The claimed results follow from Theorem 1. $\square$

REMARK 1. When the eigenvalues of $\Sigma$ are constants and do not change with $(n, p, s)$, Theorem 2 recovers a rate developed by Amini and Wainwright (2009) as a special case where Theorem 2 implies that a sufficient condition for consistent variable selection (with suitable choice of $\rho$) is $s\sqrt{\log p/n} \le c$ for a constant $c$ [according to (11) and (13)]. Amini and Wainwright (2009) also obtain a sharper sufficient condition $s \log p/n \le c'$, by assuming that the solution is rank 1. However, Krauthgamer, Nadler and Vilenchik (2013) show that, with high probability, the solution is not rank 1 unless $s\sqrt{1/n}$ is bounded by a constant.

REMARK 2. Condition (11) suggests that the required sample size needs to increase as $\lambda_1$ increases. This is because the oracle operator norm error bound of the principal subspace (i.e., assuming $J$ is known) has a factor of $\lambda_1$. In an extremal case, when $\lambda_1$ is large and $\lambda_j$ ($j \geq 2$) are much smaller, the estimation error of the leading eigenvector will likely dominate all the remaining spectral gaps, making it hard to recover the remaining eigenvectors.

**4. Agnostic inference.** Consistent estimation and variable selection inevitably depend on the existence of a "true" model. For sparse PCA, this corresponds to the assumption that the $k$-dimensional principal subspace of $\Sigma$ is (1) identifiable and (2) sparse. Under this assumption, previous work [e.g., Vu et al. (2013)] and the theory presented in Section 3 establish conditions under which consistent estimation and variable selection are possible. While these results can provide useful insights for sparse PCA and FPS, the conditions may or may not hold in practice. Therefore, it is important to understand the statistical inference problem without these assumptions. This is the *agnostic inference* perspective. Can we remove the assumptions of identifiability and sparsity? Is there an assumption-free interpretation for FPS?

Without assuming identifiability, variable selection and estimation consistency are no longer valid objectives, since there is no unique "true" parameter to estimate. For example, when $\Sigma = I$, every $k$-dimensional subspace is a principal subspace, and even if there is a unique principal subspace, it may not be sparse. To develop an assumption-free interpretation, we return to the basic objective function of PCA. Let $X$ be a random vector with covariance matrix $\Sigma$. PCA can be interpreted as a covariance maximization technique. It seeks a rank-$k$ projector $H$ that maximizes the *predictive covariance*:

$$\operatorname{trace}\bigl(\operatorname{Cov}(X, HX|H)\bigr) = \langle \Sigma, H \rangle.$$

If we interpret $H$ as a dimension-reducing transformation, then $\langle \Sigma, H \rangle$ is just the total covariance between the input $X$ and output $HX$.

4.1. *Sparse and shrinking dimension reduction.* FPS also maximizes covariance, but it replaces the rank-$k$ projector constraint on $H$ with a Fantope constraint and an additional sparsity constraint via the $(1, 1)$-norm. Let

$$\widehat{H}_R := \operatorname*{arg\,max}_{H \in \mathcal{F}^k, \|H\|_{1,1} \leq R} \langle S, H \rangle. \tag{14}$$

By Lagrangian duality, this constrained form of FPS is equivalent to the penalized form (3) in the sense that given $S$, for every $R$ there is a corresponding $\rho$ such that a solution of (3) is also a solution of (14) and vice-versa. The corresponding population version of (14) is

$$H_R := \operatorname*{arg\,max}_{H \in \mathcal{F}^k, \|H\|_{1,1} \leq R} \langle \Sigma, H \rangle. \tag{15}$$

The meaning of $H \in \mathcal{F}^k$ may be unclear since it is not necessarily a rank-$k$ projector. However, it turns out that if we regard $H$ as a linear transformation $x \mapsto Hx$, then $H$ is a *smoother matrix* [Hastie, Tibshirani and Friedman (2009), Section 5.4.1] and the Fantope coincides with a class of linear smoothers called *shrinking smoothers* [Buja, Hastie and Tibshirani (1989)]. The two essential properties of $H$ are:

1. $0 \preceq H \preceq I$. This is equivalent to the condition that

   $$\|x\|^2 \geq \|Hx\|^2 + \|x - Hx\|^2 \qquad \text{for all } x.$$

   In other words, the sum of squares of the transformation $Hx$ and its residual $x - Hx$ cannot be larger than that of $x$. A map satisfying this property is called *firmly nonexpansive*.
2. $\operatorname{trace}(H) = k$. If $H$ is a projector, then $k$ is the dimension of the projection space. It is also equal to $\operatorname{trace}[\operatorname{Cov}(\xi, H\xi)]$ when $\xi$ is a random vector with $\operatorname{Var}(\xi) = I$. By analogy, $\operatorname{trace}(H)$ is the *effective degrees of freedom* of $H$ [see Hastie, Tibshirani and Friedman (2009), Section 5.4.1].

These two properties are exactly those laid out by Hastie, Tibshirani and Friedman (2009) for smoother matrices and shrinking smoothers. In the context of dimension reduction, we call the action of $H \in \mathcal{F}^k$ *shrinking dimension reduction*.

Now we turn to the $(1, 1)$ norm constraint in (15). A natural notion of sparsity of a matrix $H \in \mathcal{F}^k$ is $\|H\|_{2,0}$, the number of nonzero rows. Here, we use the $(1, 1)$-norm as an alternative convex measure of sparsity. For $H \in \mathcal{F}^k$ we have, by Cauchy–Schwarz,

$$\|H\|_{1,1} \leq k\|H\|_{2,0}. \tag{16}$$

That is, if $\|H\|_{2,0}$ is small, then $\|H\|_{1,1}$ must also be small.

4.2. *Persistence of FPS.* Our main result in the assumption-free setting is an interpretation of the constrained form of FPS and its persistence under no assumptions on $\Sigma$.

THEOREM 3 (Persistence). *Let $X, X_1, \ldots, X_n \in \mathbb{R}^p$ be i.i.d. random vectors that satisfy the tail probability bound* (1) (*i.e., $X$ is sub-Gaussian*). *Then with probability at least $1 - 2p^{-2}$,*

$$\langle \Sigma, H_R \rangle \geq \langle \Sigma, \widehat{H}_R \rangle \geq \langle \Sigma, H_R \rangle - cR\lambda_1 \sqrt{\frac{\log p}{n}},$$

*where $c > 0$ is a constant.*

Our proof of Theorem 3 is given in Section A.2. Theorem 3 shows that the predictive covariance of FPS comes close to that of the best sparse $H$ in the Fantope. This is essentially an assumption-free interpretation.

Let

$$\Pi_{k,s} := \arg\max\{\langle \Sigma, \Pi \rangle : \Pi \text{ is a rank-}k \text{ projector and } \|\Pi\|_{2,0} \le s\}$$

be the best rank-$k$ and $s$-sparse projector. What can we say about $\widehat{H}_R$ and $\Pi_{k,s}$? In this case, (16) implies that $\|\Pi_{k,s}\|_{1,1} \le ks$. Thus $\Pi_{k,s}$ is in the feasible set of (14) if $R \ge ks$. If we do not assume any structure on $\Sigma$, Theorem 3 implies that, with high probability,

$$\langle \Sigma, \widehat{H}_R \rangle \ge \langle \Sigma, \Pi_{k,s} \rangle - cR\lambda_1 \sqrt{\frac{\log p}{n}},$$

when $R \ge ks$. If we assume in addition that $\Sigma$ does have a $k$-dimensional principal subspace involving at most $s$ variables, then the result can be strengthened to

$$\langle \Sigma, \Pi_{k,s} \rangle \ge \langle \Sigma, \widehat{H}_R \rangle \ge \langle \Sigma, \Pi_{k,s} \rangle - cR\lambda_1 \sqrt{\frac{\log p}{n}}.$$

Here, the assumption that $\Sigma$ has a sparse principal subspace is still much weaker than the sparse principal subspace condition required by the sparsistency argument in Section 3, because there is no requirement on uniqueness of the principal subspace. As a simple example, $\Sigma = I$ satisfies the sparsity condition but not the uniqueness condition.

REMARK 3 (Stability of FPS). A referee has pointed out to us that there is another interpretation of Theorem 3 in terms of the *continuity* of the maximal predictive covariance map

$$f(\Omega) := \max\{\langle \Omega, H \rangle : H \in \mathcal{F}^k, \|H\|_{1,1} \le R\}.$$

The proof of Theorem 3 implies that

$$|f(\Sigma + \Delta) - f(\Sigma)| \le 2R\|\Delta\|_{\infty,\infty}.$$

So the predictive covariance of FPS is relatively stable under perturbations of $\Sigma$ if $R\|\Delta\|_{\infty,\infty}$ is small.

**5. Discussion.** A connection between sparse PCA and sparse linear regression has been observed by Vu and Lei (2013). They established minimax rates for estimation under $\ell_2$ loss with $\ell_q$-penalized estimators with suitably defined model parameters and observed that the rates are identical to those for sparse linear regression when the effective noise variance is defined appropriately. The sparsistency result in the present paper further extends this connection to variable selection. Roughly speaking, the previously used spiked covariance model in sparse PCA, which assumes that

$$\Sigma = U\Lambda U^T + \sigma^2 I_p,$$

where $U$ is $p \times k$ orthonormal matrix and $\Lambda \succ 0$ is diagonal [see, e.g., Birnbaum et al. (2013), Cai, Ma and Wu (2013), Johnstone and Lu (2009), Ma (2013)], corresponds to the orthogonal design in linear regression, in the sense that the relevant and noise variables are not correlated. Moreover, the $\sigma^2 I$ term boosts the signal by adding $\sigma^2$ to all the relevant diagonal entries in $\Sigma$ and, therefore, thresholding based methods usually work well. The limited correlation condition developed in this paper is analogous to the irrepresentable condition [Meinshausen and Bühlmann (2006), Zhao and Yu (2006)] for $\ell_1$-penalized sparse regression (Lasso), where convex optimization methods can succeed when the correlation between relevant and noise variables is small.

When the eigenvalues of $\Sigma$ are fixed, a sufficient condition for consistent variable selection using FPS is $s \lesssim \sqrt{n/\log p}$. This is comparable to the corresponding rate developed for $k = 1$ by Amini and Wainwright (2009) when the rank of the solution is not assumed to be 1. It has been shown by Amini and Wainwright (2009) that the information-theoretic critical rate is $s \lesssim n/\log p$. That is, if $s \gg n/\log p$, no method can succeed in variable selection. It remains an open question if there exist polynomial time methods that can consistently select relevant variables in the range $\sqrt{n/\log p} \lesssim s \lesssim n/\log p$. An interesting work in this direction is that by Berthet and Rigollet (2013b), which shows that, for $k = 1$, testing a sparse PCA model in this regime is at least as hard as solving the planted clique problem beyond the well-believed computational barrier.

The predictive covariance maximization interpretation of PCA leads to a natural characterization of the Fantope as the collection of all shrinking smoothers with $k$ effective degrees of freedom. Without any assumptions on $\Sigma$, FPS gives us a dimension reducing transformation that is sparse while being computationally tractable, and it nearly approaches the best predictive covariance. In practice, it would be useful to estimate the predictive covariance of the FPS solution for a particular value of $\rho$ using risk estimates such as cross-validation. This leads to a data-driven procedure for selecting the best FPS tuning parameter $\rho$. The detailed design and properties of such a cross-validation method is an important and interesting topic for future work.

## APPENDIX: TECHNICAL PROOFS

This appendix contains detailed technical proofs. In Section A.1, we prove the deterministic sparsistency theorem (Theorem 1). Other proofs, including those of Lemma 1 and Theorem 3 are given in Section A.2.

### A.1. Proof of Theorem 1.

A.1.1. *Existence of a sparse solution.* The primal–dual witness argument starts from the dual form of the FPS problem (3). Using strong duality, we can write (3) in a equivalent min–max form:

$$\max_{H\in\mathcal{F}^k}\langle S,H\rangle-\rho\|H\|_{1,1}$$

$$\iff \max_{H\in\mathcal{F}^k}\min_{Z\in\mathbb{B}_p}\langle S,H\rangle-\rho\langle H,Z\rangle-k\rho$$

$$\iff \max_{H\in\mathcal{F}^k}\min_{Z\in\mathbb{B}_p}\langle S-\rho Z,H\rangle \tag{17}$$

$$\iff \min_{Z\in\mathbb{B}_p}\max_{H\in\mathcal{F}^k}\langle S-\rho Z,H\rangle, \tag{18}$$

where $\mathbb{B}_p=\{Z\in\mathbb{R}^{p\times p}:\operatorname{diag}(Z)=0, Z=Z^T, \|Z\|_{\infty,\infty}\le 1\}$. According to the standard Karush–Kuhn–Tucker (KKT) condition, a pair $(\widehat{H},\widehat{Z})\in\mathcal{F}^k\times\mathbb{B}_p$ is optimal for problems (17) and (18) if and only if

$$\widehat{Z}_{ij}=\operatorname{sign}(\widehat{H}_{ij})\qquad \forall i\ne j, \widehat{H}_{ij}\ne 0, \tag{19}$$

$$\widehat{Z}_{ij}\in[-1,1]\qquad \forall i\ne j, \widehat{H}_{ij}=0, \tag{20}$$

$$\widehat{H}=\operatorname*{arg\,max}_{H\in\mathcal{F}^k}\langle S-\rho\widehat{Z},H\rangle. \tag{21}$$

To proceed with the primal–dual witness argument, we first construct an additionally constrained solution $\tilde{H}$ as follows:

$$\tilde{H}=\operatorname*{arg\,max}_{H\in\mathcal{F}^k,\operatorname{supp}(\operatorname{diag}(H))\subseteq J}\langle S,H\rangle-\rho\|H\|_{1,1}. \tag{22}$$

Let $\tilde{Z}$ be a corresponding optimal dual variable. By Lemma 2, $\tilde{H}$ is a rank-$k$ projector supported on $J$.

Let $\binom{\hat{U}_J}{0}$ and $\binom{U_J}{0}$ be $p\times k$ orthogonal matrices consisting of the $k$ leading eigenvectors of $S-\rho\tilde{Z}$ and $\Sigma$, respectively, where $\hat{U}_J$ and $U_J$ are $s\times k$ orthogonal matrices. According to Lemma 2, there exists a $s\times s$ orthonormal matrix $Q$ such that $\hat{U}_J=QU_J$ and $\|Q-I\|_F\le 8\rho s/(\lambda_k-\lambda_{k+1})$.

Define a modified primal–dual pair $(\widehat{H},\widehat{Z})$ as follows (recall that $W=S-\Sigma$):

$$\widehat{H}=\tilde{H},$$

$$\widehat{Z}_{JJ}=\tilde{Z}_{JJ}, \tag{23}$$

$$\widehat{Z}_{ij}=\frac{1}{\rho}\{S_{ij}-\langle Q_{i*},\Sigma_{J,j}\rangle\},\qquad (i,j)\in J\times J^c, \tag{24}$$

$$\widehat{Z}_{ij}=\frac{1}{\rho}W_{ij},\qquad (i,j)\in(J^c)^2, i\ne j. \tag{25}$$

We need to check that $(\widehat{H}, \widehat{Z})$ is feasible for (17) and (18) and satisfies the KKT conditions (19) to (21).

*Checking feasibility*. The feasibility of $\widehat{H}$ is obvious. To check feasibility of $\widehat{Z}$, we only need to verify that $\widehat{Z}_{ij} \in [-1, 1]$ for all $(i, j) \in J \times J^c$. In fact,

$$
\begin{aligned}
|\widehat{Z}_{ij}| &\le \frac{1}{\rho}\big[|S_{ij} - \Sigma_{ij}| + \big|\Sigma_{ij} - \langle Q_{i*}, \Sigma_{J,j}\rangle\big|\big] \\
&\le \frac{1}{\rho}\big[\|W\|_{\infty,\infty} + \|(I - Q)_{i*}\| \times \|\Sigma_{J,j}\|\big] \\
&\le \frac{1}{\rho}\big[\|W\|_{\infty,\infty} + \|I - Q\|_F \|\Sigma_{J^cJ}\|_{2,\infty}\big] \\
&\le \frac{1}{\rho}\left[\|W\|_{\infty,\infty} + \frac{8\rho s}{\lambda_k - \lambda_{k+1}}\|\Sigma_{J^cJ}\|_{2,\infty}\right] \le 1,
\end{aligned}
$$

where the last inequality follows from (6).

*Checking KKT condition* (19). Because $\widehat{H}$ only has nonzero entries in $J \times J$, so $(\widehat{H}, \widehat{Z})$ satisfies (19) by construction.

*Checking KKT condition* (20). For $(i, j)$ in $J \times J$, (20) is satisfied for $(\widehat{H}, \widehat{Z})$ because the same condition is satisfied for $(\tilde{H}, \tilde{Z})$. For $(i, j) \notin J \times J$, we have $\widehat{H}_{ij} = 0$ and (20) follows from the feasibility of $\widehat{Z}$.

*Checking KKT condition* (21). Recall that $W = S - \Sigma$. Let $\tilde{W}$ be the $(p - s) \times (p - s)$ diagonal matrix that agrees with $W_{J^cJ^c}$ on diagonal entries. By Lemma 1, it suffices to show that $\binom{\hat{U}_J}{0}$ spans a $k$-dimensional principal subspace of

$$
\tilde{\Sigma} := S - \rho\widehat{Z} = \begin{pmatrix} S_{JJ} - \rho\tilde{Z}_{JJ} & Q\Sigma_{JJ^c} \\ \Sigma_{J^cJ}Q^T & \tilde{W} + \Sigma_{J^cJ^c} \end{pmatrix}, \tag{26}
$$

which is established in Lemma 3.

Now we have shown that $(\widehat{H}, \widehat{Z})$ is indeed an optimal primal–dual pair for (17) and (18), and hence $\widehat{H}$ is a solution of (3) and is also supported only on $J$.

A.1.2. *Uniqueness of solution*. Consider the elastic net version of FPS in (4) and its max–min and min–max forms using dual variable $Z \in \mathbb{B}_p := \{Z \in \mathbb{R}^{p\times p} : \operatorname{diag}(Z) = 0, Z = Z^T, \|Z\|_{\infty,\infty} \le 1\}$:

$$
\begin{aligned}
&\min_{H\in\mathcal{F}^d}\max_{Z\in\mathbb{B}_p} -\langle S, H\rangle + \rho\langle H, Z\rangle + \frac{\tau}{2}\|H\|_F^2 \\
&\qquad\Longleftrightarrow\qquad \max_{Z\in\mathbb{B}_p}\min_{H\in\mathcal{F}^k}\frac{\tau}{2}\left\|H - \frac{1}{\tau}(S - \rho Z)\right\|_F^2 - \frac{1}{2\tau}\|S - \rho Z\|_F^2.
\end{aligned}
$$

The KKT condition for optimality of $(\widehat{H}, \widehat{Z}) \in \mathcal{F}^k \times \mathbb{B}_p$ becomes

$$\widehat{Z}_{ij} = \operatorname{sign}(\widehat{H}_{ij}) \qquad \forall i \neq j, \widehat{H}_{ij} \neq 0, \tag{27}$$

$$\widehat{Z}_{ij} \in [-1, 1] \qquad \forall i \neq j, \widehat{H}_{ij} = 0, \tag{28}$$

$$\widehat{H} = \mathcal{P}_{\mathcal{F}^k}\left(\frac{1}{\tau}(S - \rho\widehat{Z})\right). \tag{29}$$

Let $\tilde{H}$, $\tilde{Z}$ be the support constrained FPS solution in (22) and $\widehat{Z}$ be the dual variable constructed in (23) to (25). We first show that $(\tilde{H}, \widehat{Z})$ is also optimal for the elastic net version of FPS when $\tau$ is small enough.

From the existence proof above and Lemma 3, we know that (i) $S - \rho\widehat{Z} = \tilde{\Sigma}$, (ii) the $k$-dimensional principal subspace of $\tilde{\Sigma}$ is spanned by $\binom{\hat{U}_J}{0}$ and (iii) $\lambda_k(\tilde{\Sigma}) - \lambda_{k+1}(\tilde{\Sigma}) > 0$.

By the construction of $\tilde{H}$, part 3 of Lemma 1 implies that when

$$0 < \tau \leq \lambda_k(\tilde{\Sigma}) - \lambda_{k+1}(\tilde{\Sigma}) \tag{30}$$

we have

$$\tilde{H} = \mathcal{P}_{\mathcal{F}^k}\left(\frac{1}{\tau}(S - \rho\widehat{Z})\right).$$

As a consequence, $(\tilde{H}, \widehat{Z})$ is also an optimal primal–dual pair for the elastic net FPS problem (4) when $\tau$ is in the range specified in (30).

Now we prove uniqueness of $\tilde{H}$ as a solution to the FPS problem (3). Assume that there is another solution $\widehat{H}' \in \mathcal{F}^k$ such that

$$\langle S, \tilde{H}\rangle - \rho\|\tilde{H}\|_{1,1} = \langle S, \widehat{H}'\rangle - \rho\|\widehat{H}'\|_{1,1}.$$

But $\tilde{H}$ is the unique solution to the elastic net FPS for $\tau > 0$ small enough, we must have $\|\widehat{H}'\|_F^2 > \|\tilde{H}\|_F^2$, and hence

$$k \geq \|\widehat{H}'\|_F^2 > \|\tilde{H}\|_F^2 = k,$$

which is a contradiction (the first inequality follows from that $\widehat{H}' \in \mathcal{F}^k$).

A.1.3. *False negative control.* The false negative control under condition (8) is obvious in view of the Frobenius norm error bound (10).

Now we prove false negative control under the entry-wise condition (9). According to Theorem 1, we know that $\widehat{H}$ is supported on $J$, and $\widehat{H}_{JJ}$ corresponds to the projector of the $k$-dimensional principal subspace of $\tilde{\Sigma}_{JJ} := S_{JJ} - \rho\widehat{Z}_{JJ}$ where $\widehat{Z}$ is the optimal dual variable.

Then it is sufficient to show that the leading eigenvector of $\tilde{\Sigma}_{JJ}$ does not have zero entries. Note that $\|\tilde{\Sigma}_{JJ} - \Sigma_{JJ}\|_{\infty,\infty} \leq 2\rho$ and the second part of assumption (9) implies that $\operatorname{sign}(\tilde{\Sigma}_{JJ}) = \operatorname{sign}(\Sigma_{JJ})$.

By the first part of assumption (9), we have $\mathrm{sign}(\tilde{\Sigma}_{JJ}) = \mathrm{sign}(\Sigma_{JJ}) = bb^T$, where $b \in \{-1, 1\}^s$. Let $B$ be the $s \times s$ diagonal matrix such that $\mathrm{diag}(B) = b$. The matrix $B\tilde{\Sigma}_{JJ}B$ has all positive entries, and hence by the Perron–Frobenius theorem, it has a unique leading eigenvector $v_1$ whose entries are all positive. As a result, the leading eigenvector of $\tilde{\Sigma}_{JJ}$ is $Bv_1$, which does not have zero entries.

A.1.4. *Auxiliary lemmas.*

LEMMA 2. *Under the assumptions in Theorem 1, let $\tilde{H}$ be the solution to the further constrained problem (22). Then $\tilde{H}$ is rank $k$ and unique. Furthermore, there exist $s \times k$ orthonormal matrices $U_J$, $\hat{U}_J$ such that:*

1. *$\binom{U_J}{0}$ and $\binom{\hat{U}_J}{0}$ span the $k$-dimensional principal subspaces of $\Sigma$ and $S - \rho\tilde{Z}$, respectively.*
2. *There exists a $s \times s$ orthonormal matrix $Q$ such that*

$$\hat{U}_J = QU_J,$$
$$\|Q - I\|_F \le \frac{8\rho s}{\lambda_k - \lambda_{k+1}}.$$

PROOF. Consider $\tilde{\Sigma}_{JJ} := S_{JJ} - \rho\tilde{Z}_{JJ}$. We know that $\tilde{H}_{JJ}$ maximizes $\langle\tilde{\Sigma}_{JJ}, H\rangle$ over all $H \in \mathcal{F}_s^k$ (the trace-$k$ Fantope of $\mathbb{R}^{s\times s}$). We argue that $\tilde{H}_{JJ}$ is unique and has rank $k$. By condition (6) and $\tilde{Z} \in \mathbb{B}_p$, we have $\|\tilde{\Sigma}_{JJ} - \Sigma_{JJ}\|_{\infty,\infty} \le 2\rho$, and hence $\|\tilde{\Sigma}_{JJ} - \Sigma_{JJ}\|_F \le 2\rho s$. On the other hand, by the (SPS) condition it is straightforward to verify that the leading $k$ eigenvectors of $\Sigma_{JJ}$ are those of $\Sigma$ confined on $J$, and hence $\lambda_\ell(\Sigma_{JJ}) = \lambda_\ell$ for all $1 \le \ell \le k$. Furthermore, since $\Sigma_{JJ}$ is a principal submatrix of $\Sigma$, we have $\lambda_{k+1}(\Sigma_{JJ}) \le \lambda_{k+1}$. Therefore,

$$\lambda_k(\tilde{\Sigma}_{JJ}) - \lambda_{k+1}(\tilde{\Sigma}_{JJ}) \ge \lambda_k(\Sigma_{JJ}) - \lambda_{k+1}(\Sigma_{JJ}) - 4\rho s \ge \lambda_k - \lambda_{k+1} - 4\rho s > 0$$

by condition (7). The first claim follows from part 1 of Lemma 1.

The second claim is trivial when $s = k$. Now we focus on the case $s > k$. By the (SPS) condition we know that the unique $k$-dimensional principal subspace of $\Sigma$ is spanned by $\binom{U_J}{0}$ where $U_J$ is a $s \times k$ orthonormal matrix. Then $U_J$ spans the $k$-dimensional principal subspace of $\Sigma_{JJ}$.

Using the fact that $\lambda_k(\Sigma_{JJ}) - \lambda_{k+1}(\Sigma_{JJ}) \ge \lambda_k - \lambda_{k+1}$, and applying Proposition 2.2 in Vu and Lei (2013), we can choose the right rotations for the columns of $\hat{U}_J$ and $U_J$ so that

$$\|\hat{U}_J - U_J\|_F \le \|\hat{U}_J\hat{U}_J^T - U_JU_J^T\|_F.$$

Using Lemma 4.2 of Vu and Lei (2013) and Cauchy–Schwarz, we have

$$\|\hat{U}_J\hat{U}_J^T - U_JU_J^T\|_F \le \frac{2}{\lambda_k - \lambda_{k+1}}\|\tilde{\Sigma}_{JJ} - \Sigma_{JJ}\|_F \le \frac{4\rho s}{\lambda_k - \lambda_{k+1}}.$$

The above two inequalities jointly imply that

$$\|\hat{U}_J - U_J\|_F \le \frac{4\rho s}{\lambda_k - \lambda_{k+1}}.$$

Now let $\hat{V} = (\hat{U}_J, \hat{U}_J^c)$ be an $s \times s$ orthonormal matrix, and similarly $V = (U_J, U_J^c)$. One can show that, using the same argument as above, $\hat{U}_J^c$ and $U_J^c$ can be chosen such that

$$\|\hat{U}_J^c - U_J^c\|_F \le \frac{4\rho s}{\lambda_k - \lambda_{k+1}}.$$

Let $Q = \hat{V} V^T$, then $QU_J = \hat{U}_J$ and

$$\|I - Q\|_F = \|(V - \hat{V})V^T\|_F = \|\hat{V} - V\|_F \le \frac{8\rho s}{\lambda_k - \lambda_{k+1}}. \qquad \square$$

LEMMA 3. *Under the assumptions of Theorem 1, let $(\tilde{H}, \tilde{Z})$ be the optimal primal–dual pair of the additionally constrained FPS problem (22). Let $\binom{\hat{U}_J}{0}$, $\binom{U_J}{0}$, and $Q$ be defined as in Lemma 2. Let $\tilde{\Sigma}$ be defined as in (26). Then*

$$\lambda_k(\tilde{\Sigma}) - \lambda_{k+1}(\tilde{\Sigma}) > 0$$

*and $\tilde{H}$ is the unique projector of the $k$-dimensional principal subspace of $\tilde{\Sigma}$.*

PROOF. We start from a decomposition of $\tilde{\Sigma}$ as follows:

$$\begin{aligned}
\tilde{\Sigma} &= \begin{pmatrix} S_{JJ} - \rho\tilde{Z}_{JJ} & Q\Sigma_{JJ^c} \\ \Sigma_{J^cJ}Q^T & \tilde{W} + \Sigma_{J^cJ^c} \end{pmatrix} \\
&= \begin{pmatrix} S_{JJ} - \rho\tilde{Z}_{JJ} - Q\Sigma_{JJ}Q^T + Q\Sigma_{JJ}Q^T & Q\Sigma_{JJ^c} \\ \Sigma_{J^cJ}Q^T & \tilde{W} + \Sigma_{J^cJ^c} \end{pmatrix} \\
&= \begin{pmatrix} S_{JJ} - \rho\tilde{Z}_{JJ} - Q\Sigma_{JJ}Q^T & 0 \\ 0 & \tilde{W} \end{pmatrix} + \begin{pmatrix} Q\Sigma_{JJ}Q^T & Q\Sigma_{JJ^c} \\ \Sigma_{J^cJ}Q^T & \Sigma_{J^cJ^c} \end{pmatrix} \\
&= \text{“noise”} + \text{“signal.”}
\end{aligned} \tag{31}$$

It can be directly verified that $\binom{\hat{U}_J}{0}$ spans the $d$-principal subspace of

$$\begin{pmatrix} Q\Sigma_{JJ}Q^T & Q\Sigma_{JJ^c} \\ \Sigma_{J^cJ}Q^T & \Sigma_{J^cJ^c} \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} \times \Sigma \times \begin{pmatrix} Q^T & 0 \\ 0 & I \end{pmatrix}. \tag{32}$$

Moreover, (32) implies that the eigenvalues of the “signal” part in the decomposition (31) are the same as those of $\Sigma$.

To sum up, we have so far shown that $\binom{\hat{U}_J}{0}$ spans the $k$-dimensional principal subspace of the signal part, with spectral gap $\lambda_k - \lambda_{k+1}$.

Next, we need to show that the $k$-dimensional principal subspace remains unchanged after adding the “noise” part.

First, the block-diagonal structure of the "noise" matrix in (31) ensures that $\binom{\hat{U}_J}{0}$ spans one of its $k$-dimensional spectral subspace (a $k$-dimensional spectral subspace of a $p \times p$ symmetric matrix $A$ means that if $v$ is in this subspace, then $Av$ is also in this subspace).

Second, we show that twice the operator norm of the "noise" part is smaller than the gap between $k$th and $(k+1)$th eigenvalues of $\tilde{\Sigma}$, which is $\lambda_k - \lambda_{k+1}$. In fact, the operator norm of the noise part does not exceed

$$\begin{aligned}\|S_{JJ} - \rho\tilde{Z}_{JJ} - \Sigma_{JJ}\| + \|\Sigma_{JJ} - Q\Sigma_{JJ}Q^T\| &\le 2\rho s + 2\|\Sigma_{JJ}\| \times \|Q - I\| \\ &\le 2\rho s + 2\lambda_1 \times 8\rho s/(\lambda_k - \lambda_{k+1}),\end{aligned}$$

where the bound on $\|Q - I\|$ comes from Lemma 2. We also have $\|\tilde{W}\| \le \rho$, which is contained within the above bound.

Therefore, by standard perturbation theory such as Weyl's inequality, the subspace spanned by $\binom{\hat{U}_J}{0}$ is the $k$-dimensional principal subspace of $\tilde{\Sigma}$ as long as

$$4\rho s + 16\sqrt{2}\lambda_1\rho s/(\lambda_k - \lambda_{k+1}) \le \lambda_k - \lambda_{k+1}, \tag{33}$$

which means that twice the noise operator norm does not exceed the spectral gap in the signal part.

When the inequality in (33) is strict, as stated in condition (7), we know that the $k$-dimensional principal subspace of $\tilde{\Sigma}$ is unique. □

### A.2. Other proofs.

PROOF OF LEMMA 1. (1) See Overton and Womersley (1992). (2) is Lemma 4.1 of Vu et al. (2013). (3) We have

$$\langle A, H\rangle - \frac{\tau}{2}\|H\|_F^2 = -\frac{\tau}{2}\|H - \tau^{-1}A\|_F^2 + \frac{1}{2\tau}\|A\|_F^2.$$

This is maximized over $H \in \mathcal{F}^k$ by $H = \mathcal{P}_{\mathcal{F}^k}(\tau^{-1}A)$. Note that by assumption $\gamma_k/\tau \ge 1$ and $\gamma_{k+1} < \gamma_k$. Then the claim follows by applying (1) and (2). □

PROOF OF THEOREM 3. Let $H_R$ be any solution of

$$\max_{H \in \mathcal{F}^k, \|H\|_{1,1} \le R} \langle \Sigma, H\rangle.$$

Then $0 \le \langle -\Sigma, \widehat{H} - H_R\rangle$, and (14) implies $0 \le \langle S, \widehat{H} - H_R\rangle$. Combining these two inequalities with the Hölder and triangle inequalities yields

$$0 \le \langle \Sigma, H_R\rangle - \langle \Sigma, \widehat{H}\rangle \le \langle S - \Sigma, \widehat{H} - H_R\rangle \le 2R\|S - \Sigma\|_{\infty,\infty}.$$

Finally, invoke (2) to complete the proof. □

**Acknowledgments.** We thank the Editors and referees for their helpful comments.

## REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the likelihood principle. In *Proceedings of the Second International Symposium of Information Theory*. Akadémiai Kiado, Budapest.

AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

BERK, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** 51–58; Correction, Ibid **37** 745–746. MR0189176

BERTHET, Q. and RIGOLLET, P. (2013a). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

BERTHET, Q. and RIGOLLET, P. (2013b). Computational lower bounds for sparse PCA. Preprint. Available at arXiv:1304.0828.

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Faund. Trends Mach. Learn.* **3** 1–122.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458

D'ASPREMONT, A., BACH, F. and EL GHAOUI, L. (2008). Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.* **9** 1269–1294. MR2426043

D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448 (electronic). MR2353806

DESHPANDE, Y. and MONTANARI, A. (2013). Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. Preprint. Available at arXiv:1304.7047.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 498–520.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I: Statistics* 221–233. Univ. California Press, Berkeley. MR0216620

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

JOLLIFFE, I. T., TRENDAFILOV, N. T. and UDDIN, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.* **12** 531–547. MR2002634

JOURNÉE, M., NESTEROV, Y., RICHTÁRIK, P. and SEPULCHRE, R. (2010). Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.* **11** 517–553. MR2600619

KEARNS, M. J., SCHAPIRE, R. E. and SELLIE, L. M. (1994). Toward efficient agnostic learning. *Mach. Learn.* **17** 115–141.

KRAUTHGAMER, R., NADLER, B. and VILENCHIK, D. (2013). Do semidefinite relaxations solve sparse PCA up to the information limit? Preprint. Available at ArXiv:1306.3690.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LOUNICI, K. (2013). Sparse principal component analysis with missing observations. *Progr. Probab.* **66** 327–356.

MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

MACKEY, L. W. (2009). Deflation methods for sparse PCA. In *Advances in Neural Information Processing Systems* 21 (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 1017–1024. Curran Associates, Red Hook, NY.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

OVERTON, M. L. and WOMERSLEY, R. S. (1992). On the sum of the largest eigenvalues of a symmetric matrix. *SIAM J. Matrix Anal. Appl.* **13** 41–45. MR1146651

PAUL, D. and JOHNSTONE, I. M. (2012). Augmented sparse principal component analysis for high dimensional data. Preprint. Available at arXiv:1202.1242.

PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2** 559–572.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

VU, V. Q. and LEI, J. (2012). Minimax rates of estimation for sparse PCA in high dimensions. In *Proc. Fifteenth International Conference on Artificial Intelligence and Statistics JMLR W&CP* **22** 1278–1286.

VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452

VU, V. Q., CHO, J., LEI, J. and ROHE, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems* (*NIPS*) 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 2670–2678. Curran Associates, Red Hook, NY.

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

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

YUAN, X.-T. and ZHANG, T. (2013). Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.* **14** 899–925. MR3063614

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jinglei@andrew.cmu.edu

DEPARTMENT OF STATISTICS
THE OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210
USA
E-MAIL: vqv@stat.osu.edu