Do Algorithms and Barriers for Sparse Principal Component Analysis Extend to Other Structured Settings?

Guanyi Wang, Mengqi Lou, and Ashwin Pananjady

***Abstract*—We study a principal component analysis problem under the spiked Wishart model in which the structure in the signal is captured by a class of union-of-subspace models. This general class includes vanilla sparse PCA as well as its variants with graph sparsity. With the goal of studying these problems under a unified statistical and computational lens, we establish fundamental limits that depend on the geometry of the problem instance, and show that a natural projected power method exhibits local convergence to the statistically near-optimal neighborhood of the solution. We complement these results with end-to-end analyses of two important special cases given by path and tree sparsity in a general basis, showing initialization methods and matching evidence of computational hardness. Overall, our results indicate that several of the phenomena observed for vanilla sparse PCA extend in a natural fashion to its structured counterparts.**

***Index Terms*—Principal component analysis, structured sparsity, nonconvex iterative optimization, computational hardness.**

I. Introduction

PRINCIPAL component analysis (PCA) is a preponderant tool for dimensionality reduction and feature extraction. PCA and its generalizations have been used for numerous applications including wavelet decomposition [5], [41], representative stock selection from business sectors [3], human face recognition [28], [45], eigen-gene selection and shaving [1], [25], [29], handwriting classification [30], clustering of functional connectivity [26], and single-cell RNA sequencing analysis [53], to name but a few.

Given a set of n samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, PCA is traditionally phrased as the problem of recovering the direction of maximal variance. However, in high dimensions when $d \gg n$, it is well-known that the vanilla estimator given by the maximal eigenvector of the sample covariance matrix of the data is inconsistent (see, e.g., Johnstone and Lu [32] and references therein). This inconsistency motivates imposing sparsity assumptions on the "ground-truth" principal component and studying the resulting problem under a generative model for the data. Indeed, sparsity has emerged as a key structural assumption inspired by the diverse applications mentioned earlier, and a wealth of literature now exists on the sparse PCA problem. In practice, additional structure exists on the ground truth principle component. For instance, in applications involving wavelet decompositions, the signal is well-modeled by structured sparsity defined on a binary tree [5]. Similarly, path sparsity on the principal component is a reasonable assumption when dealing with data representing stocks across distinct business sectors [3].

In this paper, we study a class of union-of-linearly-structured models (see Section II), which includes vanilla sparse PCA and path/tree sparse PCA as special cases. Our goal is to understand, through a statistical and computational lens, if and to what extent the theoretical results and insights developed for vanilla sparsity extend to these structured settings. In particular, given that the vanilla sparse PCA has a delicate statistical-computational gap, a conceptual question that motivates our research is

Does such a statistical-computational gap persist when additional structure is imposed in PCA?

Generally speaking, statistical-computational gaps in related problems are delicate[1], and so understanding the influence of additional structure in such problems is an important goal. In making progress toward this goal, we carry out a detailed statistical and computational study of a broad family of structured PCA problems.

Manuscript received 10 December 2023; revised 19 June 2024; accepted 24 June 2024. Date of publication 2 July 2024; date of current version 17 July 2024. The work of Guanyi Wang was supported by the National University of Singapore under AcRF Tier-1 under Grant A-8000607-00-00 22-5539-A0001. The work of Mengqi Lou and Ashwin Pananjady were supported in part by the NSF under Grant CCF-2107455 and Grant DMS-2210734 and in part by the research awards/gifts from Adobe, Amazon, and Mathworks. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Laurent Condat. *(Corresponding author: Guanyi Wang.)*

Guanyi Wang is with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576 (e-mail: guanyi.w@nus.edu.sg).

Mengqi Lou is with H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: mlou30@gatech.edu).

Ashwin Pananjady is with H. Milton Stewart School of Industrial and Systems Engineering, Atlanta, GA 30332 USA, and also with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: ashwinpm@gatech.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSP.2024.3421618, provided by the authors.

Digital Object Identifier 10.1109/TSP.2024.3421618

[1]For one such example, note that gaps disappear in the sparse stochastic block model in the presence of a "monotone adversary" [42].

A. Contributions and Organization

In Section II, we formally introduce a family of *union of linearly structured* PCA problems under the spiked Wishart model. Section III presents our main results: We begin by studying the fundamental limits of estimation under this model, providing both upper and lower bounds on the ℓ_2 error of estimation that depend on the geometry of the problem. Our upper bound is achieved by an exhaustive search algorithm, and we analyze a natural projected power method to approximately compute its solution. We show that this iterative method enjoys local geometric convergence to within a neighborhood of the ground truth solution that attains the optimal statistical rate as a function of the sample size and geometry of the problem instance. We also present a general initialization algorithm for this method. In Section IV, we study two prototypical examples of structured PCA—those given by path and tree sparsity—in an end-to-end fashion, additionally providing explicit initialization methods and evidence of computational hardness (see in particular Propositions 1 and 2). Detailed proofs of our results can be found in the supplementary material.

Through our statistical, algorithmic, and reduction-based results, we find that several features of vanilla sparse PCA—on both the statistical and computational fronts—persist and extend in natural ways to its structured counterparts. In particular, while the imposition of structure can help mildly, it does not seem to make the problem significantly easier to solve in a computationally efficient manner.

B. Related Work

Structured PCA has been studied extensively over the past two decades, and we cannot hope to cover this vast literature here. We discuss the papers most relevant to our results.

a) Optimization Algorithms for Sparse PCA: The most commonly used and studied structural assumption in PCA is (vanilla) sparsity, in which the true principal component is assumed to be k-sparse. Letting $\widehat{\mathbf{\Sigma}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$ denote the sample covariance matrix, such a sparse principal component can be found by solving the following optimization problem:

$$\max_{\boldsymbol{v}\in\mathbb{R}^d} \boldsymbol{v}^\top \widehat{\mathbf{\Sigma}} \boldsymbol{v} \quad \text{s.t.} \quad \|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_0 \le k, \tag{1}$$

where $\|\cdot\|_0$ denotes the ℓ_0 norm or number of nonzeros. This program was first proposed by [12]. In contrast to classical PCA (which is akin to program (1) but without the ℓ_0 norm constraint), solving the sparse PCA problem (1) is NP-hard. Many computationally efficient reformulations of sparse PCA have been proposed over the years. [33] give the first computational tractable method—termed SCoTLASS—which reformulates the program (1) using an ℓ_1-norm regularization akin to the LASSO [44]. [62] and [63] propose an ElasticNet version of SPCA, and [54] study connections between SCoTLASS and ElasticNet SPCA. [16], [24] propose alternative formulations and show the convergence of their alternating gradient methods to stationary points. Another approach focuses on convex relaxations of sparse PCA. For example, [17], [23], [35], [48], [61] consider a convex relaxation by lifting the variable space $\boldsymbol{v} \in \mathbb{R}^d$ to its product space, and relax to a semidefinite programming problem. More recently, [20], [21], [36] provide a more computationally scalable type of convex relaxation for problem (1) using mixed-integer programming with theoretical worst-case guarantees. Other than methods based on convex relaxation, there is also a substantial literature on specialized iterative algorithms for finding good feasible solutions. Examples include the deflation method [40], generalized power method [34], truncated power method [59], and iterative thresholding [39].

b) Statistical and Computational Limits of Sparse PCA: Several papers have established (by now classical) minimax lower bounds for sparse PCA in a purely statistical sense, i.e., without computational considerations. Examples for vector recovery in ℓ_2 norm include [8] and [14]; the latter is phrased in terms of estimating the principal subspace and considers a more general model than the rank-1 model. [49] present nonasymptotic lower and upper bounds for the minimax risk considering both row-sparse and column-sparse principal subspaces. [2] study the rank-1 spiked covariance model considered here, but establish minimax lower bounds for support recovery.

Sparse PCA has also been a key cog in the study of *computational* lower bounds in high dimensional statistics problems, and has received a lot of attention from the perspective of reductions, sum-of-squares and low-degree lower bounds, as well as approaches rooted in statistical physics; let us cover a non-exhaustive list of examples here. Assuming the planted clique conjecture, [6] show that a sub-Gaussian variant of sparse PCA is hard, in that the optimal rate of estimation is not achievable in polynomial time. Ma and Wigderson [38] show degree-4 sum of squares lower bounds for k-sparse PCA (see Section IV-B). [60] study the fundamental statistical-computational barriers of inference and estimation problems as phase transitions and develop new algorithms using techniques from statistical physics. [52] show computational lower bounds for estimation for a distributionally-robust variant of sparse PCA. [27] show computational lower bounds for sparse PCA in the spiked covariance model, and [11] provide an alternative reduction based on random rotations to strengthen these lower bounds. Ding et al. [22] explore subexponential-time algorithms for sparse PCA, and give rigorous evidence that their proposed algorithm is optimal by analyzing the low-degree likelihood ratio. [9] give a reduction from planted clique that yields the first complete characterization of the computational barrier in the spiked covariance model, providing tight lower bounds at all sparsities k.

c) Structured PCA and Related Problems: While vanilla sparsity (and the resulting sparse PCA problem) is by far the most well-studied, there also exist other examples of structure one could impose. Examples from the literature on sparse linear regression include graph sparsity [31], group sparsity structure [58], block and tree sparsity [5], and subspace constraints [7]. For PCA in particular, several structural constraints have been studied, such as non-negative orthant cone structure [43], and general cone structure [19], [56]. Asteris et al. [3] study path-sparse structure in the PCA problem. Some of these papers study fundamental limits of estimation for their specific forms of structure, and Asteris et al. [3] and [56] propose specialized projected power methods. [13] present a unified framework for

the statistical analysis of structured principal subspace estimation and lower and upper bounds on the minimax risk. In recent work, [37] study structured PCA under the assumption that the true principal component is generated from an L-Lipschitz continuous generative model, showing that the projected power method enjoys local geometric convergence. While their result is related in spirit to a subset of our results on the projected power method, our structural assumptions are different (see Definition 1 and discussions following Theorem 2) for a detailed comparison. There are also papers that study computational hardness in structured settings, both from the perspective of low-degree polynomials [4] and reductions from the so-called "secret-leakage" variant of the planted clique conjecture [10]. Our work adds to this literature for a particular family of structured PCA problems.

II. Problem Setting, Background, and Examples

Throughout this paper, we operate under the spiked Wishart model. Assume that our data set consists of n i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^n$ drawn from a d-dimensional Gaussian distribution with zero-mean and covariance $\boldsymbol{\Sigma} := \lambda \boldsymbol{v}_* \boldsymbol{v}_*^\top + \boldsymbol{I}_{d\times d}$. For brevity, we use $\mathcal{D}(\lambda; \boldsymbol{v}_*) := \mathcal{N}(\boldsymbol{0}_d, \lambda \boldsymbol{v}_* \boldsymbol{v}_*^\top + \boldsymbol{I}_{d\times d})$ to denote the distribution of each $\boldsymbol{x}_i$. Here $\lambda > 0$ represents the *strength of the signal*, and $\boldsymbol{v}_*$ is a d-dimensional, unit-norm *ground truth* vector that we wish to estimate. In addition to the unit norm condition, we also assume the inclusion $\boldsymbol{v}^* \in \mathcal{M}$, where $\mathcal{M}$ is a known union of subspaces satisfying a certain *union of linear structures* assumption defined below.

Definition 1: **Union of linear structures condition.** Let $\mathcal{B} := \{\phi_1, \ldots, \phi_d\}$ be an orthonormal basis of $\mathbb{R}^d$ and $\mathcal{L} := \{L_1, \ldots, L_M\}$ be a collection of M distinct linear subspaces such that for each $m \in [M] := \{1, \ldots, M\}$, we have $L_m = \text{span}(\mathcal{B}_m)$ for some $\mathcal{B}_m \subseteq \mathcal{B}$. We say set $\mathcal{M}$ obeys the union of linear structures condition if $\mathcal{M} := \bigcup_{m=1}^M L_m$, i.e., $\mathcal{M}$ is the union of all linear subspaces in $\mathcal{L}$.

Remark 1: It is worth noting that the union of linear structures condition in Definition 1 resembles a structured sparsity condition. Indeed, using the rotation invariance of the Gaussian distribution, the problem of estimating $\boldsymbol{v}_*$ from observations $\{\boldsymbol{x}_i\}_{i=1}^n$ is *statistically* equivalent to estimating the structured-sparse vector $\boldsymbol{\Phi}^\top \boldsymbol{v}_*$ from $\{\boldsymbol{\Phi}^\top \boldsymbol{x}_i\}_{i=1}^n$, where $\boldsymbol{\Phi} \in \mathbb{R}^{d\times d}$ is an orthonormal matrix with columns $\phi_1, \ldots, \phi_d$. However, the two problems may not be *computationally* equivalent when $\boldsymbol{\Phi}$ is unknown. Here, we provide an example (see Example 1 in Appendix A) to illustrate that if an efficient projection oracle onto the union of subspaces $\mathcal{M}$ is accessible, then it is more computationally efficient to estimate the vector $\boldsymbol{v}_*$ directly, rather than to estimate $\boldsymbol{\Phi}^\top \boldsymbol{v}_*$ from $\{\boldsymbol{\Phi}^\top \boldsymbol{x}_i\}_{i=1}^n$ by first computing $\boldsymbol{\Phi}$. Accordingly, the rest of the paper assumes that $\boldsymbol{\Phi}$ is unknown, and that we have access to a projection oracle onto the union of subspaces $\mathcal{M}$.

A. Examples of Union of Linearly Structure in Section II

Clearly, vanilla sparse PCA is covered by our formulation. We instantiate the union of linear structures assumption with two other canonical examples.

1) Example 1: Tree-Sparse PCA: Motivated by applications in signal and image processing and computer graphics [5], a particular model for the underlying signal is *tree sparsity* in an underlying basis. In particular, consider the following simplified model for tree-sparsity with one-dimensional signals and binary wavelet trees as a typical such instance. We require some notation to introduce it formally.

Given a natural number h, a *complete binary tree* or CBT of size $d = 2^h - 1$ is given by the following construction. Create h levels $\{1, \ldots, h\}$, with $2^{\ell-1}$ nodes in ℓ-th level. Index each node from 1 to d, top to bottom and left to right in the following way. The root node r_{CBT} of CBT has index 1, and for any node with index $i \in \{2, \ldots, 2^{h-1} - 1\}$, its parent is the node with index $\lfloor \frac{i}{2} \rfloor$ and its children are the nodes with indices $2i, 2i+1$. Define the collection of vertex sets

$$\mathcal{T}^k := \big\{T : |T| = k, \text{ root node } 1 \in T, \text{ the subgraph of CBT induced by } T \text{ is connected}\big\}.$$

Abusing notation slightly, consider a bijection between the coordinates of any d-dimensional vector and the vertices of a CBT. The vector $\boldsymbol{v}_*$ is said to be *k-tree-sparse* if $\text{supp}(\boldsymbol{v}_*) \in \mathcal{T}^k$.

Therefore, tree-sparse PCA is a specific example of union of linear structures in our formulation. To see this, let $\boldsymbol{e}_i \in \mathcal{S}^{d-1}$ denote the i-th standard basis vector in $\mathbb{R}^d$, and set

$$\mathcal{B} := \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}, \text{ and } \mathcal{L} := \big\{L = \text{span}(\{\boldsymbol{e}_i\}_{i\in T}) \mid T \in \mathcal{T}^k\big\}$$

in Definition 1.

2) Example 2: Path-Sparse PCA: Another commonly used variant of union-of-linearly structured PCA is path-sparse PCA [3], in which the support set of $\boldsymbol{v}_*$ forms a path on an underlying directed acyclic graph $G = (V, E)$. For a vertex v in this graph, let $\delta_{\text{out}}(v)$ denote the out-neighborhood of v.

Definition 2: (d, k)-Layered Graph. A directed acyclic graph $G = (V, E)$ is a (d, k)-layered graph if

- $V = \{v_s, v_t\} \cup \widetilde{V}$ such that $|\widetilde{V}| = d - 2$ and $v_s, v_t \notin \widetilde{V}$.
- $\widetilde{V} = \cup_{i=1}^k V_i$ where $V_i \cap V_j = \emptyset$ for all $i \neq j \in [k]$ and $|V_1| = \cdots = |V_k| = \frac{d-2}{k}$.
- $\delta_{\text{out}}(v) = V_{i+1}$ for all $v \in V_i$ and $i = 1, \ldots, k-1$, and
- $\delta_{\text{out}}(v_s) = V_1$ and $\delta_{\text{out}}(v) = \{v_t\}$ for all $v \in V_k$.

Let $G = (V, E)$ be a (d, k)-layered graph and we define the collection of vertex sets

$$\mathcal{P}^k := \big\{P \subseteq V \mid v_s, v_t \in P \text{ and } |P \cap V_i| = 1 \ \forall\ i \in [k]\big\}.$$

Once again, we consider the natural bijection between the coordinates of any d-dimensional vector and the vertices of a (d, k)-layered graph, and a vector $\boldsymbol{v}_*$ is said to be *k-path-sparse* if $\text{supp}(\boldsymbol{v}_*) \in \mathcal{P}^k$. It is straightforward to see that the set of all *k-path-sparse* vectors satisfies the union of linear structures condition in Definition 1 with

$$\mathcal{B} := \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}, \text{ and } \mathcal{L} := \{L = \text{span}(\{\boldsymbol{e}_i\}_{i\in P}) \mid P \in \mathcal{P}^k\}.$$

B. Notation

We use $\boldsymbol{I}_{d\times d}$ to denote the d-by-d identity matrix, and $\lambda_i(\boldsymbol{M})$ to denote the i-th largest eigenvalue of a symmetric

matrix $\boldsymbol{M}$. We use $\boldsymbol{X} := [\boldsymbol{x}_1 \mid \cdots \mid \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ to denote the sample matrix where the i-th row of $\boldsymbol{X}$ is the i-th sample $\boldsymbol{x}_i$. The sample covariance matrix is given by $\widehat{\boldsymbol{\Sigma}} := \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$, and we let

$$\boldsymbol{W} := \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \tag{2}$$

denote the $d \times d$ matrix of noise. For any linear subspace $L \subseteq \mathbb{R}^d$ and its projection matrix $\boldsymbol{P}_L \in \mathbb{R}^{d\times d}$, we use $\widehat{\boldsymbol{\Sigma}}_L := \boldsymbol{P}_L^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{P}_L$ to denote the sample covariance matrix restricted to the subspace L. We also use the analogous notation $\boldsymbol{\Sigma}_L := \boldsymbol{P}_L^\top\boldsymbol{\Sigma}\boldsymbol{P}_L$ and $\boldsymbol{W}_L := \boldsymbol{P}_L^\top\boldsymbol{W}\boldsymbol{P}_L$. We index the subspaces $L_1, \ldots, L_M$ in some consistent lexicographic order. We reserve the notation $\mathcal{M} := \bigcup_{m=1}^M L_m$ to denote the set containing $\boldsymbol{v}_*$, and the notation $L_* \in \{L_1, \ldots, L_M\}$ to denote the specific linear subspace that contains $\boldsymbol{v}_*$, with ties broken lexicographically. We use $[M] := \{1, \ldots, M\}$ to denote the index set indexed from 1 to M. We let $\mathcal{S}^{d-1} := \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$ denote the unit ℓ_2-sphere in d-dimensional Euclidean space. For any subspace L, let $\widehat{\boldsymbol{v}}_L := \operatorname{argmax}_{\boldsymbol{v}\in\mathcal{S}^{d-1}} \boldsymbol{v}^\top\widehat{\boldsymbol{\Sigma}}_L\boldsymbol{v} = \operatorname{argmax}_{\boldsymbol{v}\in\mathcal{S}^{d-1}\cap L} \boldsymbol{v}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{v}$ be the leading eigenvector of the restricted sample covariance $\widehat{\boldsymbol{\Sigma}}_L$. For an arbitrary symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ and set $S \subseteq \mathbb{R}^d$, define the scalar

$$\rho(\boldsymbol{M}, S) := \max_{\|\boldsymbol{v}\|_2=1, \boldsymbol{v}\in S} \left|\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{v}\right|. \tag{3}$$

For two sequences of non-negative reals $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$, we use $f_n \gtrsim g_n$ to indicate that there is a universal positive constant C such that $f_n \leq Cg_n$ for all $n \geq 1$. We also use standard order notation $f_n = O(g_n)$ to indicate that $f_n \lesssim g_n$ and $f_n = \tilde{O}(g_n)$ to indicate that $f_n \lesssim g_n \ln^c n$ for some universal constant c. We say that $f_n = \Omega(g_n)$ (resp. $f_n = \tilde{\Omega}(g_n)$) if $g_n = \Omega(f_n)$ (resp. $g_n = \Omega(f_n)$). We use $f_n = \Theta(g_n)$ (resp. $f_n = \tilde{\Theta}(g_n)$) if $f_n = O(g_n)$ and $f_n = \Omega(g_n)$ (resp. $f_n = \tilde{O}(g_n)$ and $f_n = \tilde{\Omega}(g_n)$). We say that $f_n = o(g_n)$ (resp. $f_n = \tilde{o}(g_n)$) when $\lim_{n\to\infty} f_n/g_n = 0$ (resp. $\lim_{n\to\infty} f_n/(g_n \ln^c n) = 0$ for some universal constant c). We also use $f_n = \omega(g_n)$ to indicate that $\lim_{n\to\infty} f_n/g_n = \infty$. Throughout, we use $c, c_1, c_2, \ldots$ and $C, C_1, C_2, \ldots$ to denote universal positive constants, and their values may change from line to line.

III. General Results

In this section, we present our general results for union of linearly structured PCA, covering both fundamental limits of estimation and local convergence properties of a projected power method. Recall the notation $\rho(\boldsymbol{M}, S)$ from Eq. (3) for any symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ and set $S \subseteq \mathbb{R}^d$. We let

$$\widehat{\boldsymbol{v}}_{\mathsf{ES}} := \operatorname*{argmax}_{\boldsymbol{v}\in\mathcal{M}\cap\mathcal{S}^{d-1}} \boldsymbol{v}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{v} \tag{4}$$

denote the general exhaustive search estimator.

A. Fundamental Limits of estimation

We begin by studying the fundamental limits of estimation for linearly structured PCA, without computational considerations. These serve as baselines for the results to follow. We first introduce some notation before presenting main results. Recall $\mathcal{L} = \{L_1, \ldots, L_M\}$, the collection of M linear subspaces, and subsets of bases $\mathcal{B}_m \subseteq \mathcal{B} = \{\phi_1, \ldots, \phi_d\}$ such that $L_m = \operatorname{span}(\mathcal{B}_m)$. For each $m \in [M]$, define the characteristic vector $\boldsymbol{z}_m \in \{0,1\}^d$ of each subset $\mathcal{B}_m$ as follows

$$\boldsymbol{z}_m(i) := \begin{cases} 1 & \text{if} \quad \phi_i \in \mathcal{B}_m \\ 0 & \text{if} \quad \phi_i \notin \mathcal{B}_m \end{cases}, \quad \text{for all } i \in [d], \tag{5}$$

where $\boldsymbol{z}_m(i)$ is the i-th entry of $\boldsymbol{z}_m$. We further define

$$i_* := \operatorname*{argmax}_{i\in[d]} \sum_{m=1}^M \boldsymbol{z}_m(i) \tag{6}$$

as the index with the most ones among $\{\boldsymbol{z}_m\}_{m=1}^M$, breaking ties lexicographically. In words, this is the index of the basis vector that appears in the most subspaces. Now let

$$\mathcal{Z}_* := \{\boldsymbol{z}_m \in \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M\} \mid \boldsymbol{z}_m(i_*) = 1\}. \tag{7}$$

be the set of characteristic vectors with $\boldsymbol{z}_m(i_*) = 1$. For any fixed integer $r \geq 0$ and characteristic vector $\boldsymbol{z} \in \{\boldsymbol{z}_m\}_{m=1}^M$, we use $\mathcal{N}_H(\boldsymbol{z}; r) := \{\boldsymbol{z}' \in \mathcal{Z}_* \mid \delta_H(\boldsymbol{z}, \boldsymbol{z}') \leq r\}$ to denote the neighborhood of $\boldsymbol{z}$ in $\mathcal{Z}_*$ with Hamming ball distance $\delta_H(\boldsymbol{z}, \boldsymbol{z}') := |\{i : \boldsymbol{z}(i) \neq \boldsymbol{z}'(i)\}|$ at most r. We further state Assumption 1 for the minimax lower bound.

Assumption 1: This assumption has two parts:

(a) For all $m \in [M]$, $|\mathcal{B}_m| = k$ for some $k \leq d$.

(b) There exists $\xi \in [3/4, 1)$ such that

$$\frac{|\mathcal{Z}_*|}{\max_{\boldsymbol{z}\in\mathcal{Z}_*} |\mathcal{N}_H(\boldsymbol{z}; 2(1-\xi)k)|} \geq 16. \tag{8}$$

Assumption 1(a) is clearly satisfied by vanilla sparse PCA, tree-sparse PCA, and path-sparse PCA. For a general $\mathcal{L}$, one can always set $k = \max_{m\in[M]} |\mathcal{B}_m|$. Assumption 1(b), on the other hand, controls the ratio of the sizes between the largest neighborhood $\mathcal{N}_H(\boldsymbol{z}; 2(1-\xi)k)$ (among $\boldsymbol{z} \in \mathcal{Z}_*$) and $\mathcal{Z}_*$. Geometric intuition for this assumption will be provided shortly. It is worth noting that the specific constant 16 in Ineq. (8) is arbitrary, and any constant greater than 2 can be used. We choose 16 for simplicity and convenience in presenting the subsequent theoretical results (Theorem 1(b)).

We are now poised to state the main result of this subsection. Recall that L_* denotes the subspace containing the vector $\boldsymbol{v}_*$. Let $\widehat{L} \in \mathcal{L}$ be the linear subspace such that $\widehat{\boldsymbol{v}}_{\mathsf{ES}} \in \widehat{L}$ (once again breaking ties lexicographically) and let $\widehat{F} := \mathsf{conv}(\widehat{L} \cup L_*)$.

Theorem 1: Suppose the union-of-linear structures condition in Definition 1 holds.

(a) Let $\widehat{\boldsymbol{v}}_{\mathsf{ES}}$ be defined in equation (4). Without loss of generality, suppose $\langle \boldsymbol{v}_*, \widehat{\boldsymbol{v}}_{\mathsf{ES}} \rangle \geq 0$. Then for all $\boldsymbol{v}_* \in \mathcal{S}^{d-1} \cap \mathcal{M}$, we have

$$\|\widehat{\boldsymbol{v}}_{\mathsf{ES}} - \boldsymbol{v}_*\|_2 \leq \frac{2\sqrt{2}}{\lambda}\rho\big(\boldsymbol{W}, \widehat{F}\big), \tag{9a}$$

where the function ρ is defined in Eq. (3).

(b) Let $\xi \in [3/4, 1)$ such that Assumption 1 holds. We have the minimax lower bound

$$\inf_{\widehat{\boldsymbol{v}}} \sup_{\boldsymbol{v}_* \in \mathcal{S}^{d-1} \cap \mathcal{M}} \mathbb{E}\left[\left\|\widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top - \boldsymbol{v}_*\boldsymbol{v}_*^\top\right\|_F\right] \geq \frac{\sqrt{2(1-\xi)}}{4} \cdot \min\left\{1, \sqrt{\frac{1+\lambda}{8\lambda^2}} \cdot \sqrt{\log\left(\frac{|\mathcal{Z}_*|}{\max_{\boldsymbol{z}\in\mathcal{Z}_*} |\mathcal{N}_H(\boldsymbol{z}; 2(1-\xi)k)|}\right)}\right\}. \tag{9b}$$

Here, the infimum is taken over all measurable functions of the observations $\{\boldsymbol{x}_i\}_{i=1}^n$, which are drawn i.i.d. from the distribution $\mathcal{D}(\lambda; \boldsymbol{v}_*)$.

Theorem 1(a) provides a deterministic upper bound on the ℓ_2 error between the estimate $\widehat{\boldsymbol{v}}_{\mathsf{ES}}$ and the ground truth $\boldsymbol{v}_*$, showing that this error can be bounded on the order $\rho(\boldsymbol{W}, \widehat{F})$ for any fixed λ. We provide the proof of this result in Section B1 of the supplementary material. While the result is deterministic, we will see that Eq. (9a) nearly matches the minimax lower bound Eq. (9b) for our special cases of interest. Consequently, we use Theorem 1(a) as a heuristic baseline to assess the performance of efficient algorithms.

On its own, Theorem 1(b) provides a minimax lower bound that depends on the local structure of $\mathcal{M}$ around any choice of ground truth $\boldsymbol{v}_*$. The proof uses the generalized Fano inequality [46], and we construct a rich packing set $\mathcal{V}_\epsilon$ in $\mathcal{S}^{d-1} \cap \mathcal{M}$ (i.e., $\mathcal{V}$ in Proposition 4 of Section B2 in the supplementary material) such that the points in $\mathcal{V}_\epsilon$ are $\mathcal{O}(\epsilon)$ separated in some appropriate distance measure. In contrast to existing proofs for sparse PCA [47] and path PCA [3], the set $\mathcal{V}_\epsilon$ here is constructed so that there exists a common support index (i.e., the index i_*, defined in Eq (6)) for every point $\boldsymbol{v} \in \mathcal{V}_\epsilon$ that one can use to construct the packing. On a related note, a paper by Cai et al. [13] studies the minimax risk of a general structured principal subspace estimation problem, including vanilla sparse PCA as a special case. These bounds are phrased in terms of critical inequalities that arise from local packing numbers (see [50], [55]). Our lower bound instead takes a more global approach, which we show suffices for union-of-linear structure. In particular, the minimax lower bound (9b) is controlled by the relative ratio between $|\mathcal{N}_H(\boldsymbol{z}; 2(1-\xi)k)|$ and $|\mathcal{Z}_*|$: Our assumption in Ineq. (8) avoids the scenario that many linear subspaces heavily overlap on a few bases.

Finally, it is instructive to note that Theorem 1(b) recovers the existing minimax lower bound for vanilla sparse PCA [Theorem 2.1, 47]. Indeed, supposing that $d \gg k$ and applying Theorem 1(b) for sparse PCA, we obtain the known lower bound

$$\inf_{\widehat{\boldsymbol{v}}} \sup_{\boldsymbol{v}_* \in \mathcal{S}^{d-1}, \|\boldsymbol{v}_*\|_0 \leq k} \mathbb{E}\left[\left\|\widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top - \boldsymbol{v}_*\boldsymbol{v}_*^\top\right\|_F\right] \gtrsim \min\left\{1, \sqrt{\frac{1+\lambda}{8\lambda^2}}\sqrt{\frac{k\log d}{n}}\right\}. \tag{10}$$

The proof of inequality (10) is provided in Section A.6.6 of Wang et al. [51] for completeness due to page limit.

Algorithm 1 Projected Power Method

Input: Sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$.

1: **Initialize** with a vector $\boldsymbol{v}_0 \in \mathcal{M} \cap \mathcal{S}^{d-1}$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Compute $\tilde{\boldsymbol{v}}_{t+1} = \widehat{\boldsymbol{\Sigma}}\boldsymbol{v}_t / \left\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{v}_t\right\|_2$.
4: Project $\boldsymbol{v}_{t+1}^{\mathcal{M}} = \Pi_{\mathcal{M}}(\tilde{\boldsymbol{v}}_{t+1})$.
5: Normalize to unit sphere $\boldsymbol{v}_{t+1} = \frac{\boldsymbol{v}_{t+1}^{\mathcal{M}}}{\|\boldsymbol{v}_{t+1}^{\mathcal{M}}\|_2} \in \mathcal{M} \cap \mathcal{S}^{d-1}$.
6: **end for**

Output: $\boldsymbol{v}_T$.

In Section IV, we provide novel corollaries for tree-sparse PCA and path-sparse PCA.

B. A Locally Convergent Projected Power Method

In Section III-A, we studied the fundamental limits of the problem, where our upper bounds were achieved by the exhaustive search estimator $\widehat{\boldsymbol{v}}_{\mathsf{ES}}$. Given the computational challenge of searching over every linear subspace $L \in \mathcal{L}$, we propose the following iterative projected power method (Algorithm 1) and show that with access to a suitable exact projection oracle, it locally converges to a statistical neighborhood of the ground truth.

Definition 3 (Exact projection): For all $\boldsymbol{v} \in \mathbb{R}^d$, let

$$\Pi_{\mathcal{M}}(\boldsymbol{v}) := \operatorname*{argmin}_{\boldsymbol{v}' \in \mathcal{M}} \|\boldsymbol{v}' - \boldsymbol{v}\|_2 = \operatorname*{argmin}_{\boldsymbol{v}' \in L_m, m \in [M]} \|\boldsymbol{v}' - \boldsymbol{v}\|_2,$$

where ties between subspaces are broken lexicographically.

Owing to the tie-breaking rule, this projection is always unique. As we will see in Section IV, an exact projection oracle $\Pi_{\mathcal{M}}$ can be constructed efficiently (in time nearly logarithmic in M) in some specific examples of union-of-linearly structured PCA. We are now in a position to present the projected power method, described formally in Algorithm 1.

Using the notation $\rho(\boldsymbol{M}, S)$ from Eq. (3), we define $F^* := \operatorname{argmax}_F \rho(\boldsymbol{W}, F)$ s.t. $F = \operatorname{conv}(L_{m_1} \cup L_{m_2} \cup L_{m_3})$, $\forall\, m_1, m_2, m_3 \in [M]$. We now state the definition of a "good region"; Theorem 2 to follow shows that once Algorithm 1 is initialized in this region, it will converge geometrically to a neighborhood of the ground truth $\boldsymbol{v}_*$.

Definition 4 (Good region): For eigen gap $\lambda > 2\rho(\boldsymbol{W}, F^*)$, we define the good region

$$\mathbb{G}(\lambda) = \left\{\boldsymbol{v} \in \mathcal{M} \cap \mathcal{S}^{d-1} : \langle \boldsymbol{v}, \boldsymbol{v}_* \rangle \geq t_1(\lambda)\right\}, \quad \text{where}$$
$$t_1(\lambda) := \frac{4}{\lambda + 1 - \rho(\boldsymbol{W}, F^*)} + \frac{5\rho(\boldsymbol{W}, F^*)}{\lambda - 2\rho(\boldsymbol{W}, F^*)}.$$

Note that $\mathbb{G}(\lambda)$ becomes a larger set as λ increases. To ensure that such a good region is non-empty, it is necessary to have $t_1(\lambda) < 1$. In our proof of convergence (see Supplementary Material C), we require that the good region is not just

non-empty but large enough. In particular, we require the eigengap λ to be large enough so that

$$t_2(\lambda) := \frac{4}{\lambda + 1 - \rho(\boldsymbol{W}, F^*)} + \frac{10\rho(\boldsymbol{W}, F^*)}{\lambda - 2\rho(\boldsymbol{W}, F^*)} < 1. \quad (11)$$

Note that this automatically ensures that $t_1(\lambda) < 1$ since $t_2(\lambda) > t_1(\lambda)$. We are now poised to state our main result for this subsection.

Theorem 2: Suppose the eigengap satisfies $\lambda > 2\rho(\boldsymbol{W}, F^*)$, and condition (11) holds. Suppose in Algorithm 4 the initialization satisfies $\boldsymbol{v}_0 \in \mathbb{G}(\lambda)$. Then for all $t \geq 1$, we have

$$\|\boldsymbol{v}_{t+1} - \boldsymbol{v}_*\|_2 \leq \frac{1}{2^t} \cdot \|\boldsymbol{v}_0 - \boldsymbol{v}_*\|_2 + \frac{6\rho(\boldsymbol{W}, F^*)}{\lambda - 2\rho(\boldsymbol{W}, F^*)}. \quad (12)$$

The proof of Theorem 2 can be found in Section C. Once (a) an exact projection oracle $\Pi_{\mathcal{M}}$ is accessible; and (b) an initial vector $\boldsymbol{v}_0$ is in the good region $\mathbb{G}(\lambda)$, Theorem 2 ensures a deterministic convergence result. Note that the result parallels that of [59] for vanilla sparse PCA, where $\rho(\boldsymbol{W}, F^*) = O(\sqrt{k \log d/n})$. The key additional technique that we use to control the error accumulated at each iteration is based on an "equivalent replacement" step; see Section C2.

The projected power method was also recently analyzed by Liu et al. [37] for PCA with generative models when given access to an exact projection oracle. While they also proved local geometric convergence results given access to a sufficiently correlated initialization, there are significant differences in the assumptions of that paper and our own. First, our work imposes the union-of-linear structure assumption on the principal component, which is an altogether different structural assumption from a generative model. Given this, our proof techniques differ significantly from those of Liu et al. [37]. Second, we present a computationally efficient initialization method and matching evidence of computational hardness for two prototypical examples; see below.

C. Initialization Method

Recall that Theorem 2 requires an initialization $\boldsymbol{v}_0$ in the good region $\mathbb{G}(\lambda)$. In this subsection, we provide such an initialization method (see Algorithm 2) that works when given a projection oracle, provided the following assumption holds.

Assumption 2: The set $\mathcal{M}$ satisfies

$$\mathcal{M} \subseteq \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_0 = k\}, \quad \text{where} \quad k \in \mathbb{N}.$$

Assumption 2 is not guaranteed by Definition 1, but includes many typical examples. For instance, the sets $\mathcal{T}^k$ and $\mathcal{P}^k$ for tree-sparse or path-sparse PCA, respectively, satisfy Assumption 2 in addition to union-of-linear structure. Moreover, if the orthonormal matrix $\boldsymbol{\Phi}$ is known, one can reformulate the problem as estimating the structured-sparse vector $\boldsymbol{\Phi}^\top \boldsymbol{v}_*$ from observations $\{\boldsymbol{\Phi}^\top \boldsymbol{x}_i\}_{i=1}^n$ (see Remark 1).

Theorem 3: Suppose Assumption 2 holds and $k^2 \leq d/e$. There exists a tuple of universal, positive constants (C_1, C_2, C_3, C) such that the following holds. Suppose $n \geq \max\{C \log d, k^2\}$ and let $\tau_* := C_1 \max\{\lambda, 1\}\sqrt{\log(d/k^2)}$. Set the thresholding level according to

$$\tau := \begin{cases} \tau_* & \text{when } \tau_* \leq \sqrt{\log d}/2, \\ C_2\tau_* & \text{when } \tau_* \geq \sqrt{\log d}/2, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

Then for any $0 < c_0 < 1$, if

$$n \geq n_0(c_0) := \frac{18C_3 \max\{\lambda^2,\ 1\}k^2}{2(1 - c_0)^2\lambda^2} \log(d/k^2),$$

then the initial vector $\boldsymbol{v}_0 \in \mathcal{S}^{d-1} \cap \mathcal{M}$ obtained from Algorithm 2 satisfies $\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle \geq c_0$ with probability $1 - C' \exp(-\min\{\sqrt{d}, n\}/C')$ for some positive constant C'.

The proof of Theorem 3, which builds on existing results in [18], can be found in Section D. Let us now show that the output of this algorithm serves as a valid initialization for the projected power method, since this is not immediate given that the event $\mathcal{E}_1 = \{\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle \geq c_0\}$ depends on the samples $\{\boldsymbol{x}_i\}_{i=1}^n$. Recall the quantities $t_1(\lambda)$ and $t_2(\lambda)$ in Definition 4 and Eq. (11), respectively. In Theorem 3, set $c_0 := t_2(\lambda)$ and recall that $t_2(\lambda) > t_2(\lambda)$ by definition. Suppose $\lambda \geq 5$ for convenience. Then it can be shown that the event $\mathcal{E}_2 = \{t_1(\lambda) < t_2(\lambda) = c_0 < 7/8\} \subseteq \{\rho(\boldsymbol{W}, F^*) < 9/400\}$ occurs with probability at least $1 - C' \exp(-\min\{\sqrt{d}, n\}/C')$. Consequently, on the high probability event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have that the initialization $\boldsymbol{v}_0$ obtained by Algorithm 2 satisfies $\boldsymbol{v}_0 \in \mathbb{G}(\lambda)$. The projected power method can thus be employed after this initialization to guarantee convergence to a small neighborhood of $\boldsymbol{v}_*$.

A key feature of Theorem 3 is the lower bound $n_0 = \Theta(k^2 \log(d/k^2))$ on the number of samples required for the Algorithm 3 to succeed. Note that this is of a strictly larger order than the number of samples required information-theoretically

Algorithm 2 Initialization Method – Covariance Thresholding with Projection Oracle

Input. $\{\boldsymbol{x}_i\}_{i=1}^n$, parameter $k \in \mathbb{N}$, thresholding parameter τ and exact projection $\Pi_{\mathcal{M}}$.

1: Compute covariance matrix $\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top / n$.

2: Set the soft-thresholding matrix $\widehat{\boldsymbol{G}}(\tau)$ as:

$$\begin{aligned} &\text{If} \quad \widehat{\boldsymbol{\Sigma}}_{ij} - [\boldsymbol{I}_d]_{ij} \geq \tau/\sqrt{n}, \\ &\qquad \text{then} \quad [\widehat{\boldsymbol{G}}(\tau)]_{ij} = \widehat{\boldsymbol{\Sigma}}_{ij} - [\boldsymbol{I}_d]_{ij} - \tau/\sqrt{n}; \\ &\text{else if} \quad \widehat{\boldsymbol{\Sigma}}_{ij} - [\boldsymbol{I}_d]_{ij} \leq -\tau/\sqrt{n}, \\ &\qquad \text{then} \quad [\widehat{\boldsymbol{G}}(\tau)]_{ij} = \widehat{\boldsymbol{\Sigma}}_{ij} - [\boldsymbol{I}_d]_{ij} + \tau/\sqrt{n}; \\ &\text{else} \quad [\widehat{\boldsymbol{G}}(\tau)]_{ij} = 0. \end{aligned}$$

3: Compute $\widehat{\boldsymbol{v}}_{\text{soft}} := \max_{\|\boldsymbol{v}\|_2 = 1} \boldsymbol{v}^\top \widehat{\boldsymbol{G}}(\tau) \boldsymbol{v}$ as the leading eigenvector of $\widehat{\boldsymbol{G}}(\tau)$.

4: Project $\boldsymbol{v}_0 := \Pi_{\mathcal{M}}(\widehat{\boldsymbol{v}}_{\text{soft}}) / \|\Pi_{\mathcal{M}}(\widehat{\boldsymbol{v}}_{\text{soft}})\|_2$.

Return $\boldsymbol{v}_0 \in \mathcal{S}^{d-1} \cap \mathcal{M}$.

Algorithm 3 Exact Projection Oracle – Path Sparse PCA

Input: A (d,k)-layered graph G, a vector $\boldsymbol{v} \in \mathbb{R}^d$.

1: **for** $\ell = 1, \ldots, k$ **do**
2: Pick S_ℓ the index set of the ℓ-th layer in G.
3: Compute path sparsity vector $\boldsymbol{v}^{\mathsf{PS}}$ as follows: for its sub-vector $\boldsymbol{v}^{\mathsf{PS}}_{S_\ell}$, set

$$[\boldsymbol{v}^{\mathsf{PS}}_{S_\ell}]_i := \begin{cases} [\boldsymbol{v}_{S_\ell}]_i & \text{if component } i \text{ has} \\ & \text{the largest absolute value,} \\ & \text{breaking ties lexicographically} \\ 0 & \text{otherwise} \end{cases}.$$

4: **end for**
5: Normalize $\boldsymbol{v}^{\mathsf{PS}} := \boldsymbol{v}^{\mathsf{PS}}/\|\boldsymbol{v}^{\mathsf{PS}}\|_2$.

Output: $\boldsymbol{v}^{\mathsf{PS}}$.

even for vanilla sparse PCA—this is a well-known phenomenon. In the next section, we show that even with the additional structure afforded by tree and path sparsity, this larger sample size is in some sense necessary for computationally efficient algorithms.

IV. End-to-End Analysis for Specific Examples

In this section, we provide end-to-end analyses for path-sparse and tree-sparse PCA, including results on their information-theoretic limits of estimation as well as the performance of the projected power method when initialized using covariance thresholding. We complement these with what may be considered as the main results of this section: matching suggestions of computational hardness.

A. Path-Sparse PCA

1) Fundamental Limits for Path-Sparse PCA: Recall the notation $\mathcal{P}^k$ as the structure set of path-sparse PCA from Section II-A2. We write $\boldsymbol{v} \in \mathcal{P}^k$ if the support set satisfies $\mathsf{supp}(\boldsymbol{v}) \in \mathcal{P}^k$. We use

$$\widehat{\boldsymbol{v}}_{\mathsf{PS}} := \operatorname*{argmax}_{\boldsymbol{v}} \ \boldsymbol{v}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{v} \ \text{ s.t. } \ \boldsymbol{v} \in \mathcal{S}^{d-1} \cap \mathcal{P}^k \tag{14}$$

to denote the corresponding estimate from exhaustive search.

Corollary 1: There exists a pair of positive constants (c, C) such that the following holds.

(a) Without loss of generality, assume $\langle \boldsymbol{v}_*, \widehat{\boldsymbol{v}}_{\mathsf{PS}} \rangle \geq 0$. Then for any $c_1 > 0$ and $\boldsymbol{v}_* \in \mathcal{S}^{d-1} \cap \mathcal{P}^k$, we have

$$\left\| \widehat{\boldsymbol{v}}_{\mathsf{PS}} - \boldsymbol{v}_* \right\|_2 \leq C \left(\frac{1+\lambda}{\lambda} \right) \sqrt{\frac{3(\ln d - \ln k)k + c_1 k}{n}}$$

with probability at least $1 - 2\exp(-c_1 k)$.

(b) Suppose that $d \geq 16k^2$ and $k \geq 4$. Then we have the minimax lower bound

$$\inf_{\widehat{\boldsymbol{v}}} \sup_{\boldsymbol{v}_* \in \mathcal{S}^{d-1} \cap \mathcal{P}^k} \mathbb{E}\left[\left\| \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top - \boldsymbol{v}_*\boldsymbol{v}_*^\top \right\|_F \right] \geq c \cdot \min\left\{ 1, \sqrt{\frac{1+\lambda}{8\lambda^2}} \sqrt{\frac{k \cdot \left(\frac{\ln d}{2} - \ln k\right)}{n}} \right\}.$$

Here, the infimum is taken over all measurable functions of the observations $\{\boldsymbol{x}_i\}_{i=1}^n$ drawn i.i.d. from the distribution $\mathcal{D}(\lambda; \boldsymbol{v}_*)$.

Corollary 1(a) gives an upper bound on the estimation error of $\widehat{\boldsymbol{v}}_{\mathsf{PS}}$ by showing that the statistical noise term[2] $\rho(\boldsymbol{W}, P^*)$ is of the order $(\lambda+1)\sqrt{k \cdot (\ln d - \ln k)/n}$. The minimax lower bound obtained in Corollary 1(b) is of the same order as the minimax lower bound given in [Theorem 1, 3] with the outer degree parameter $|\Gamma_{\text{out}}(v)| = (d-2)/k$. The full proof of Corollary 1 is omitted due to space constraints, and can be found in [Section A.6.1, Wang et al. 51].

As we can observe from Fig. 1, methods with path-sparse projection outperform the methods with k-sparse projection with respect to the performance metric point distance and probability of success, especially as dimension d and sparsity level k increases.

2) Local Convergence and Initialization:

a) Exact Projection Oracle: We build the exact projection oracle for path-sparse PCA $\Pi_{\mathcal{P}^k}$ by picking the component with the largest absolute value in each partition (layer) for a given (d,k)-layered graph G. The formal procedure is given in Algorithm 3 as follows, and has running time $O(d)$.

Corollary 2: Suppose the initialization $\boldsymbol{v}_0$ in Algorithm 1 satisfies $\boldsymbol{v}_0 \in \mathcal{P}^k \cap \mathcal{S}^{d-1}$ and $\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle \geq 1/2$. There exists a tuple of universal positive constants (c, C_1, C_2, C_3) such that for $\lambda \geq C_1$, $n \geq C_2 k \ln(d)$, and all $t \geq 1$, the iterate $\boldsymbol{v}_t$ from Algorithm 1 satisfies

$$\|\boldsymbol{v}_t - \boldsymbol{v}_*\|_2 \leq \frac{1}{2^t} \cdot \|\boldsymbol{v}_0 - \boldsymbol{v}_*\|_2 + C_3 \sqrt{\frac{k(2\ln d - \ln k)}{n}},$$

with probability at least $1 - \exp(-ck)$.

Corollary 2 is proved by applying Theorem 2, and the full proof can be found in [Section A.6.2, Wang et al. 51].

The final problem is to obtain an initialization $\boldsymbol{v}_0$. To do so, note that the set $\mathcal{P}^k$ satisfies Assumption 2, leading to the following corollary of Theorem 3.

Corollary 3: Assume $k^2 \leq d/e$. There exists a pair of universal positive constants (C, C') such that if $n \geq \max\{C \log d, k^2\}$ and $n \geq C' \max\{1, \lambda^{-2}\} \log(d/k^2) k^2$, then the initial vector $\boldsymbol{v}_0 \in \mathcal{S}^{d-1} \cap \mathcal{P}^k$ obtained from Algorithm 2 satisfies $\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle \geq 7/8$ with probability $1 - C' \exp(-\min\{\sqrt{d}, n\}/C')$.

In words, Corollary 3 provides an initialization method whose outputs can be used for the general projected power method (Algorithm 1) for path-sparse PCA when the number of samples satisfies[3] $n \gtrsim k^2 \log(d/k^2)$.

As previously mentioned, there is a gap between the condition $n \gtrsim k$ required for Corollary 2 and the stronger condition above. We will now show evidence that k^2 samples are necessary. In particular, we will show that no randomized

[2] As expected, this term does not differ significantly from the corresponding term for vanilla sparse PCA, since the number of sparsity patterns for path sparse PCA $|\mathcal{P}^k|$ is on the order $(d/k)^k$.

[3] The constant $7/8$ in $\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle \geq 7/8$ can be replaced by any positive constant within $(0,1)$ provided it ensures the good region condition $\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle > t_2(\lambda)$.

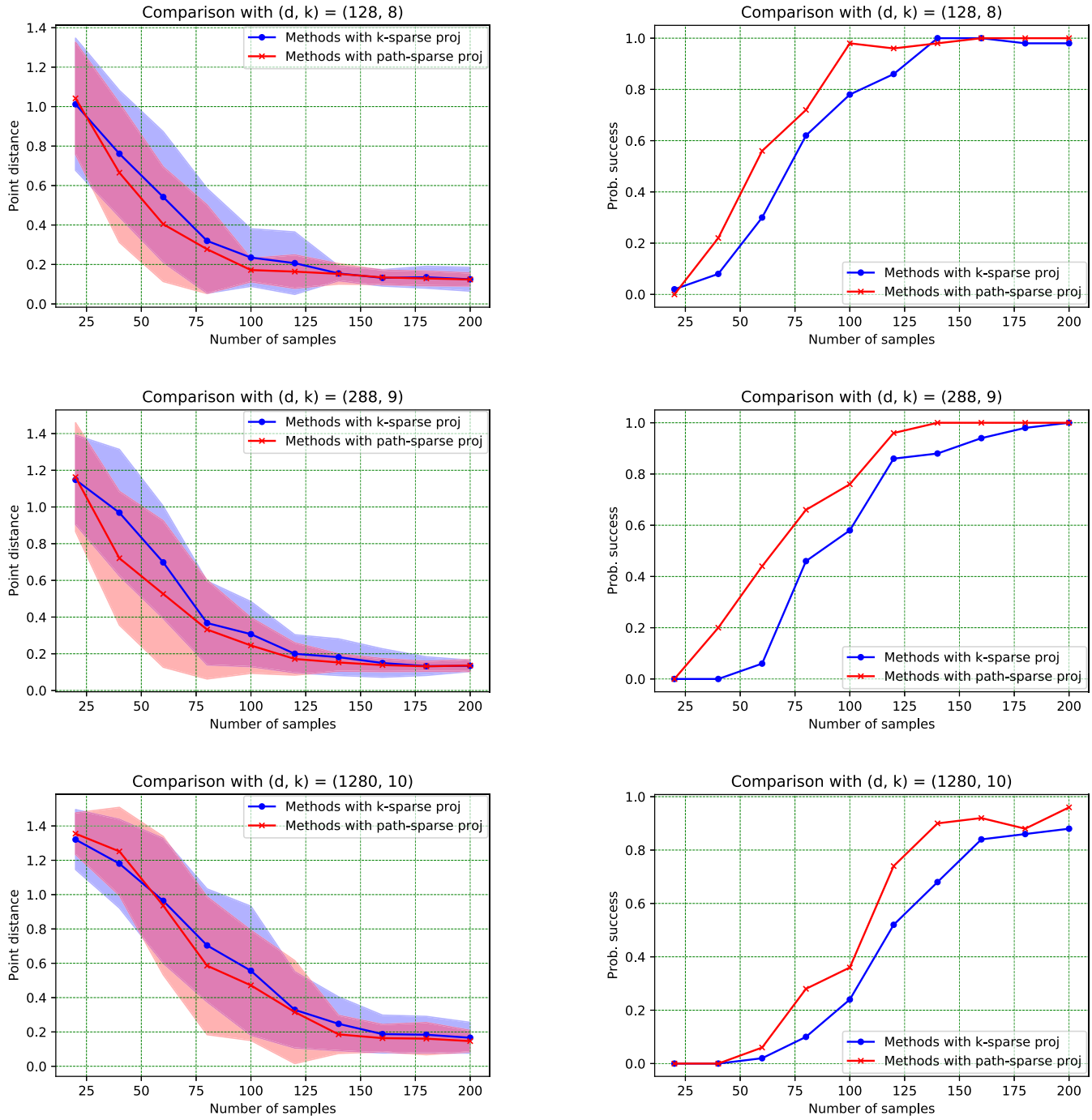

Fig. 1. Given the sparsity k, the number of nodes L in each layer, sample dimension $d = L \times k$, and eigengap λ, we choose a particular path sparsity support set $P_* \in \mathcal{P}^k$ and set the ground truth vector $\boldsymbol{v}_*$ as $[\boldsymbol{v}_*]_i = \pm\frac{1}{\sqrt{k}}$ if $i \in P_*$ and $[\boldsymbol{v}_*]_i = 0$ if $i \notin P_*$. Given a tuple of (λ, d, k, n), for each trial, we generate samples from the distribution $\mathcal{D}(\lambda, \boldsymbol{v}_*)$ based on the Wishart model in Section II, and we run Algorithm 2 (covariance thresholding) for initialization, and Algorithm 1 (projected power method) with general k-sparse projection or with path-sparse projection for local refinement. Each trial is repeated 50 times independently. We set $\lambda = 3$ and choose $(d, L, k) = (128, 16, 8), (288, 32, 9), (1280, 128, 10)$. For each choice of (d, L, k), we simulate for each $n = \{20, 40, \ldots, 200\}$. In the left column, we plot the ℓ_2 distance $\|\boldsymbol{v}_T - \boldsymbol{v}_*\|_2$ versus the number of samples n. The two curves in each panel correspond to the averaged values over 50 independent trials of the proposed methods with general k-sparse projection or with path-sparse projection; the shaded parts represent the empirical standard deviations over 50 trials. As we can observe, using path-sparse projection achieves smaller estimation error (for a given, small sample size) than using general k-sparse projection. In the right column, we further plot of the success probability of support recovery of the methods using general k-sparse projection or using path-sparse projection verse the number of samples n. The support of $\boldsymbol{v}_*$ is considered as successfully recovered if $\mathsf{supp}(\boldsymbol{v}_T) = P_*$. The success probability is then computed as the ratio of the number of trials that successfully recover the support over 50 independent trials. For a fixed small sample size, we observe that using path-sparse projection achieves higher success probability of support recovery compared with using the vanilla k-sparse projection.

polynomial-time algorithm can "solve" (i.e. produce a consistent estimate for) path-sparse when $n \ll k^2$, provided we assume the average-case hardness of the secret-leakage planted clique problem. This can be regarded as the main takeaway for path-sparse PCA: The additional structure has minimal effect on its statistical and computational limits.

3) Average-Case Hardness of Path Sparse PCA: This section focuses on the average-case hardness of the path sparse PCA, which is obtained via a reduction from the K-partite planted clique (PC) detection problem, which is in turn conjectured to be hard.

Definition 5: **Secret Leakage PC$_\mathcal{D}$ Detection Problem,** [10]. Given a distribution $\mathcal{D}$ on K-subsets of $[N]$, let $\mathcal{G}_\mathcal{D}(N, K, 1/2)$ be the distribution on N-vertex graphs sampled by first sampling $G \sim \mathcal{G}(N, 1/2)$ and $S \sim \mathcal{D}$ independently and then planting a K-clique on the vertex set S in G. The secret leakage PC$_\mathcal{D}$ detect problem PC$_\mathcal{D}(N, K, 1/2)$ is defined as the resulting hypothesis testing problem between

$$H_0 : \ G \sim \mathcal{G}(N, 1/2) \quad \text{and} \quad H_1 : \ G \sim \mathcal{G}_\mathcal{D}(N, K, 1/2).$$

Now consider the following K-partite PC as a special case of the secret leakage PC$_\mathcal{D}$ detection problem.

Definition 6: **K-Partite Planted Clique Detection Problem (with source and terminal).** The K-partite planted clique detection problem K-PC$(N, K, 1/2)$ is a special case of the secret leakage planted clique detection problem PC$_\mathcal{D}(N, K, 1/2)$. Here the vertex set of G has two special vertices: source and terminal, and the remaining vertices are evenly partition into K parts of size $(N-2)/K$. The distribution $\mathcal{D}$ always picks source, terminal and uniformly picks one element at random in each part.

Like the well-known planted clique conjecture, the K-Partite PC problem K-PC$(N, K, 1/2)$ is believed to satisfy the following hardness conjecture.

Conjecture 1: **K-Partite PC Hardness Conjecture**, restatement of [10]. Suppose that $\{\mathcal{A}_N\}$ is a sequence of randomized polynomial time algorithms $\mathcal{A}_N : \mathcal{G}_N \to \{0, 1\}$ and K_N is a sequence of positive integers satisfying that $\limsup_{N\to\infty} \log_N K_N < 1/2$ with $\mathcal{G}_N$ the set of graphs with N nodes. Then if G is an instance of K-PC$(N, K_N, 1/2)$, it holds that $\liminf_{N\to\infty} (\mathbb{P}_{H_0}[\mathcal{A}_N(G) = 1] + \mathbb{P}_{H_1}[\mathcal{A}_N(G) = 0]) \geq 1$.

Definition 7: **Qualified Estimator.** A qualified estimator $\widehat{\boldsymbol{v}}(n, d_n, k_n, \lambda_n, \epsilon)$ for path-sparse PCA is a sequence of functions $\mathtt{Est}_n : \mathbb{R}^{d_n \times n} \to \mathbb{R}^{d_n}$ mapping $\{\boldsymbol{x}_i\}_{i=1}^n \mapsto \widehat{\boldsymbol{v}}$ such that if the set of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ are drawn i.i.d. from $\mathcal{D}(\lambda_n, \boldsymbol{v}_*)$ for some $\boldsymbol{v}_* \in \mathcal{S}^{d_n-1} \cap \mathcal{P}^{k_n}$ then $\liminf_{n\to\infty} \Pr\left\{\|\widehat{\boldsymbol{v}} - \boldsymbol{v}_*\|_2 < \frac{1}{4}\right\} \geq \frac{1}{2} + \epsilon$ for some fixed $0 < \epsilon < 1/2$.

From this point onward, we do not make ϵ explicit when referring to a qualified estimator. It suffices for the reader to think of it as a small positive constant that does not depend on n. Geometrically, a qualified estimator $\widehat{\boldsymbol{v}}$ exhibits proximity to the ground truth $\boldsymbol{v}_* \in \mathcal{S}^{d_n-1} \cap \mathcal{P}^{k_n}$ with probability at least $1/2 + \epsilon$ as $n \to \infty$. Note that Definition 7 does not require explicit control on the behavior of $\widehat{\boldsymbol{v}}$ for a general vector $\boldsymbol{v}_* \notin \mathcal{S}^{d_n-1} \cap \mathcal{P}^{k_n}$.

It is also worth noting (using Corollary 2 and Corollary 3 and the corresponding algorithms) that our end-to-end estimator for path-sparse PCA is a polynomial-time computable qualified estimator provided $n \geq Ck^2 \log(d/k)$ and $\lambda = \Omega(1)$.

Proposition 1: There exists a universal constant $c > 0$ such that the following holds. Let $1/2 \leq \beta < 1$ and $0 < \epsilon < 1/2$ be fixed. Here, we use integer j as our index parameter. Suppose the sequence of parameters $\{(k_j, d_j, \lambda_j, \tau_j)\}_{j\in\mathbb{N}}$ is in the parameter regime

$$k_j = \lceil j^\beta \rceil, \quad d_j = j, \quad \lambda_j = \frac{k_j^2}{\tau_j \cdot j} \cdot \frac{(\log 2)^2}{4(6\log(j) + 2\log 2)},$$

where τ_j is an arbitrarily slowly growing function of j. If the K-Partite PC hardness conjecture (Conjecture 1) holds, then there is no qualified estimator $\widehat{\boldsymbol{v}}(n_j, d_j, k_j, \lambda_j, \epsilon)$ running in time polynomial in d_j when the sample size n_j satisfies $n_j \leq c\left(\frac{k_j^2}{2\tau_j \log k_j}\right)$.

The proof of Proposition 1 is given in Section E1 of the Supplementary Material. In particular, when the eigengap satisfies[4] $\lambda = \Theta(1)$, it shows that $n = \widetilde{\Omega}\left(k^2\right)$ is necessary for computationally efficient estimation.

B. Tree-Sparse PCA

1) Fundamental Limits for Tree-Sparse PCA: Recall the notation $\mathcal{T}^k$ as the set of all rooted binary subtrees in the underlying *complete binary tree* from Section II-A1. We write $\boldsymbol{v} \in \mathcal{T}^k$ if the support set of $\boldsymbol{v}$ satisfies $\mathsf{supp}(\boldsymbol{v}) \in \mathcal{T}^k$. Let

$$\widehat{\boldsymbol{v}}_{\mathsf{TS}} := \operatorname*{argmax}_{\boldsymbol{v}} \ \boldsymbol{v}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{v} \ \text{ s.t. } \ \boldsymbol{v} \in \mathcal{S}^{d-1} \cap \mathcal{T}^k \tag{15}$$

denote the estimator obtained from exhaustive search.

Corollary 4: There exists a pair of positive constants (c, C) such that the following holds.

(a) Without loss of generality, suppose $\langle \boldsymbol{v}_*, \widehat{\boldsymbol{v}}_{\mathsf{TS}} \rangle \geq 0$. Then for any $c_1 > 0$ and $\boldsymbol{v}_* \in \mathcal{S}^{d-1} \cap \mathcal{T}^k$, we have

$$\left\|\widehat{\boldsymbol{v}}_{\mathsf{TS}} - \boldsymbol{v}_*\right\|_2 \leq C\left(\frac{1+\lambda}{\lambda}\right) \cdot \sqrt{\frac{(3 + \ln 2 + c_1)k}{n}}$$

with probability at least $1 - 2\exp(-c_1 k)$.

(b) We have the minimax lower bound

$$\inf_{\widehat{\boldsymbol{v}}} \sup_{\boldsymbol{v}_* \in \mathcal{S}^{d-1} \cap \mathcal{T}^k} \mathbb{E}\left[\left\|\widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top - \boldsymbol{v}_*\boldsymbol{v}_*^\top\right\|_F\right] \geq c \cdot \min\left\{\frac{1}{4\sqrt{\log k}}, \ \frac{1}{4}\sqrt{\frac{1+\lambda}{8\lambda^2}}\sqrt{\frac{k/\log k}{n}}\right\}.$$

Here, the infimum is taken over all measurable functions of the observations $\{\boldsymbol{x}_i\}_{i=1}^n$ drawn i.i.d. from the distribution $\mathcal{D}(\lambda; \boldsymbol{v}_*)$.

The full proof of Corollary 4 is provided in [Section A.6.3, Wang et al. 51]. The term $\sqrt{k/n}$ arises from evaluating the cardinality of the set $\mathcal{T}^k$ in tree-sparse PCA. In particular, we have $|\mathcal{T}^k| \leq (2e)^k/(k+1)$ [5], and taking logarithms

[4]This can be ensured for dimension $d_j = j$ growing such that $\frac{k_j}{\tau_j d_j \log d_j} = \Theta(1)$.

results in a logarithmic factor gain over vanilla sparse PCA. Corollary 4(b) provides a minimax lower bound of $\Omega(\sqrt{k/(n\log k)})$ for tree-sparse PCA, which has a logarithm gap $\sqrt{1/\log k}$ compared with the upper bound in Corollary 4(a). This gap is small for small k, but we conjecture that it can be eliminated.

Remark 2: Compared with the fundamental limits for vanilla sparse PCA, the upper bounds for tree-sparse PCA in Corollary 4 save a factor $\log d$, which parallels the model-based compressed sensing literature. The saving could be significant in practice when d is large (see Fig. 2 to follow)—indeed, this is one of the successes behind model-based compressive sensing.

2) Local Convergence and Initialization:

a) Exact Projection Oracle: We use the projection method proposed in [15] as our tractable exact projection oracle $\Pi_{\mathcal{T}^k}$ for tree sparse PCA. This oracle has running time $O(kd)$. With our projection oracle in hand, we can now state our corollaries for the projected power method for tree sparse PCA.

Corollary 5: Suppose in Algorithm 1 that the initialization $\boldsymbol{v}_0 \in \mathcal{T}^k \cap \mathcal{S}^{d-1}$ satisfies $\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle \geq 1/2$. There exists a tuple of universal positive constants (c, C_1, C_2, C_3) such that for $\lambda \geq C_1$, $n \geq C_2 k$ and all $t \geq 1$, the iterate $\boldsymbol{v}_t$ from Algorithm 4 satisfies

$$\|\boldsymbol{v}_t - \boldsymbol{v}_*\|_2 \leq \frac{1}{2^t} \cdot \|\boldsymbol{v}_0 - \boldsymbol{v}_*\|_2 + C_3\sqrt{\frac{k}{n}},$$

with probability at least $1 - \exp(-ck)$.

Corollary 5 can be derived directly from Theorem 2, but we provide the full proof in [Section A.6.4, Wang et al. 51] for completeness. We can also use the exact projection oracle $\Pi_{\mathcal{T}^k}$ to obtain the following corollary for our initialization method.

Corollary 6: Assume $k^2 \leq d/e$. There exists a pair of universal positive constants (C, C') such that if $n \geq \max\{C\log d, k^2\}$ and $n \geq C' \max\{1, \lambda^{-2}\}\log(d/k^2)k^2$, then Algorithm 2 returns an initial vector $\boldsymbol{v}_0 \in \mathcal{S}^{d-1} \cap \mathcal{T}^k$ satisfying $\langle \boldsymbol{v}_0, \boldsymbol{v}_* \rangle \geq 1/2$ with probability $1 - C'\exp(-\min\{\sqrt{d}, n\}/C')$.

Like Corollary 3, it is straightforward to see that Corollary 6 follows from Theorem 3 by specifying $c_0 = 1/2$.

Corollary 6 shows that provided $n = \Omega(k^2)$, the output $\boldsymbol{v}_0 \in \mathcal{S}^{d-1} \cap \mathcal{T}^k$ satisfies the initialization condition required for the subsequent projected power method to succeed. Putting these two results together, we have produced an end-to-end and computationally efficient algorithm that produces a statistically efficient solution provided $n = \Omega(k^2)$. The next section is concerned with the question of whether the condition $n = \Omega(k^2)$ is necessary for polynomial-time algorithms.

3) SDP Hardness for Tree Sparse PCA: To understand the aforementioned gap in sample size, we now provide a computational lower bound for a class of SDP solutions to tree-sparse PCA, showing that they require on the order of k^2 samples.

To make things formal, we consider the following subclass of tree sparse PCA problems: every entry of the k tree-sparse ground truth unit vector $\boldsymbol{v}_*$ only takes one of the values $\{0, \pm k^{-1/2}\}$. With knowledge of this side information in addition to tree sparsity, the natural choice of exhaustive estimator is given by the maximizer of the following optimization problem:

$$\begin{aligned} \max_{\boldsymbol{v}} \quad & \boldsymbol{v}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{v} \\ \text{s.t.} \quad & \|\boldsymbol{v}\|_2^2 = 1,\ \|\boldsymbol{v}\|_0 = k \\ & \boldsymbol{v}(i)^2 \leq \boldsymbol{v}(\lfloor i/2 \rfloor)^2 \ \text{ for all } 2 \leq i \leq d. \end{aligned} \tag{16}$$

The natural semidefinite programming (SDP) relaxation of the program (16) is then given by

$$\begin{aligned} \mathsf{SDP}(\widehat{\boldsymbol{\Sigma}}) = \max_{\boldsymbol{M} \in \mathbb{R}^{d\times d}} \quad & \sum_{i=1}^{d}\sum_{j=1}^{d} \widehat{\boldsymbol{\Sigma}}_{ij} \boldsymbol{M}_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{d} \boldsymbol{M}_{ii}^2 = 1 \\ & \sum_{i=1}^{d}\sum_{j=1}^{d} |\boldsymbol{M}_{ij}| \leq k \\ & \boldsymbol{M} \succeq \boldsymbol{0}_{d\times d} \\ & \boldsymbol{M}_{ii} \leq \boldsymbol{M}_{\lfloor i/2 \rfloor \lfloor i/2 \rfloor} \\ & \text{for all } \ 2 \leq i \leq d. \end{aligned} \tag{17}$$

It is well-known that for vanilla sparse PCA, the SDP attains the best-known sample complexity among all polynomial time algorithms. Proving a lower bound for this class of algorithms is thus powerful—when this subclass of low-degree estimators fails at the indicated threshold, it suggests a natural hardness result.

Proposition 2: Suppose data $\boldsymbol{X}$ are drawn from the distribution $\mathcal{D}(\lambda; \boldsymbol{v}_*)$ with ground truth $\boldsymbol{v}_*$ given by a k tree-sparse unit vector with every entry of taking one of the values in the set $\{0, \pm k^{-1/2}\}$. There exists a tuple of universal positive constants (c, c_1, C, C_1) such that for $c_1 d \leq n \leq C_1 d$, $n \leq ck^2$ and $1 \leq \lambda \leq \frac{d}{Cn}$, the optimal solution $\boldsymbol{M}_*$ of the SDP relaxation (17) satisfies $\left\|\boldsymbol{M}_* - \boldsymbol{v}_*\boldsymbol{v}_*^\top\right\|_2 \geq \frac{1}{5}$ with probability at least $1 - \tilde{c}d^{-\tilde{c}}$ for some constant $\tilde{c} \geq 1$.

In words, Proposition 2 shows that unless the number of samples satisfies $n \geq C'k^2$ for some positive constant $C' \geq c$, the optimal solution $\boldsymbol{M}_*$ of the SDP relaxation (17) fails to estimate the ground truth consistently, even with the side information that its entries take only one of three values. The full proof of Proposition 2 can be found in [Section A.7.2, Wang et al. 51], and is built on the techniques proposed in [Section 4, 38].

V. Discussion

We studied the local convergence properties of the projected power method in a general class of structured PCA problems. We also established the fundamental limits of estimation in this family of problems, and studied a general family of initialization methods. Our work generalizes these statistical and algorithmic results from vanilla sparse PCA to this more general class of models. We specialized our results to two commonly used notions of structure—given by tree and path sparsity—showing end-to-end estimation algorithms accompanied by evidence of computational hardness.

Let us close with some potential questions for future investigation. The first is to generalize these results to other

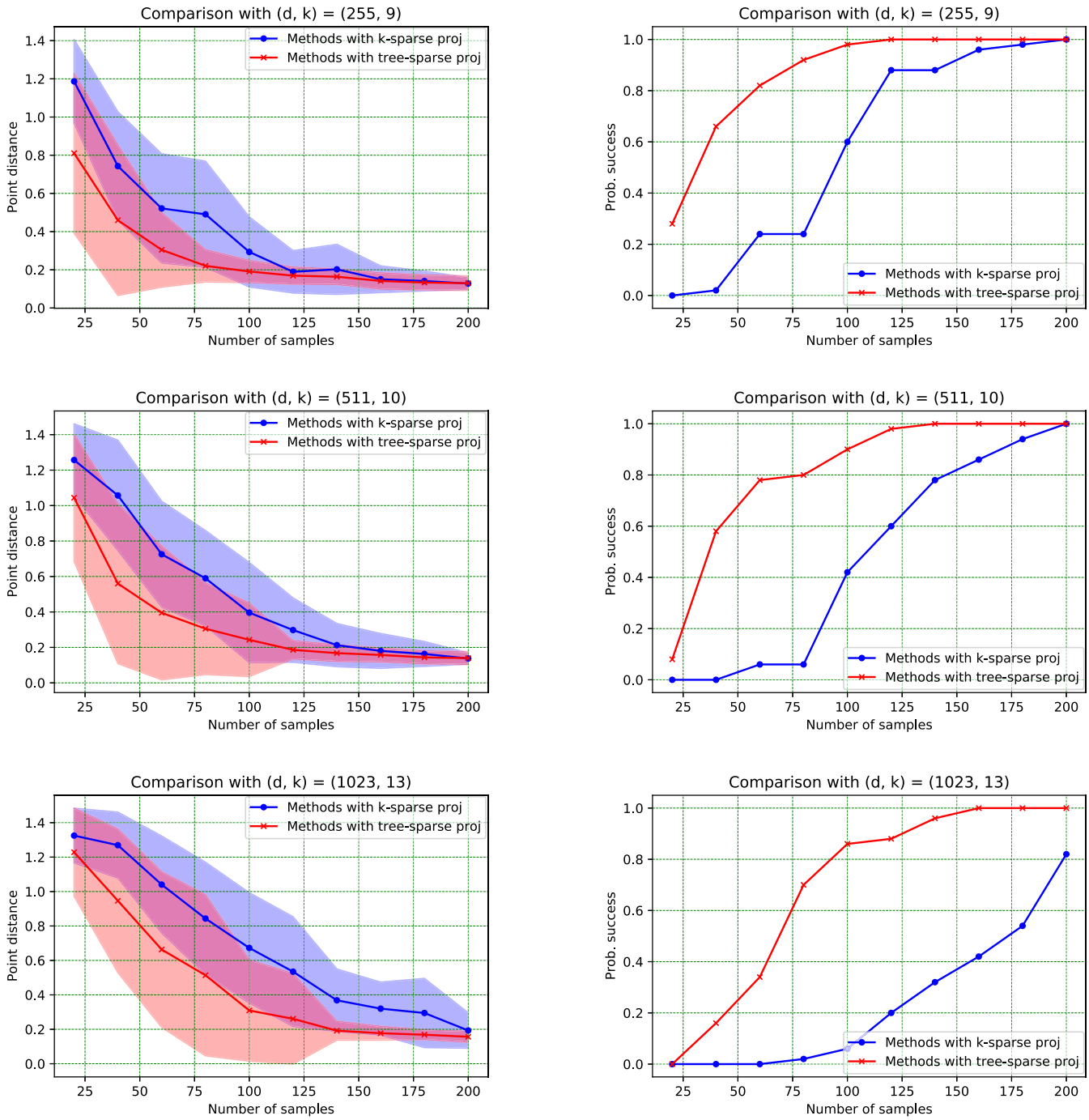

Fig. 2. Given the sample dimension $d = 2^L - 1$, sparsity k, and eigengap λ, we choose a particular tree sparsity support set $T_* \in \mathcal{T}^k$ and set the ground truth vector $\boldsymbol{v}_*$ as $[\boldsymbol{v}_*]_i = \pm\frac{1}{\sqrt{k}}$ if $i \in T_*$ and $[\boldsymbol{v}_*]_i = 0$ if $i \notin T_*$. Given a tuple of (λ, d, k, n), for each trial, we generate samples from the distribution $\mathcal{D}(\lambda, \boldsymbol{v}_*)$ based on the Wishart model in Section II, and we run Algorithm 2 for initialization, and Algorithm 1 with general k-sparse projection or with tree-sparse projection for local refinement. Each trial is repeated 50 times independently. We set $\lambda = 3$ and choose $(d, k) = (255, 9), (511, 10), (1023, 13)$. For each choice of (d, k), we simulate for each $n = \{20, 40, \ldots, 200\}$. In the left column, we plot the ℓ_2 distance $\|\boldsymbol{v}_T - \boldsymbol{v}_*\|_2$ versus the number of samples n. The two curves in each panel correspond to the averaged values over 50 independent trials of the proposed methods with general k-sparse projection or with tree-sparse projection; the shaded parts represent the empirical standard deviations over 50 trials. As we can observe, using tree-sparse projection achieves smaller estimation error (for a given, small sample size) than using general k-sparse projection. In the right column, we further plot of the success probability of support recovery of the methods using general k-sparse projection or using tree-sparse projection verse the number of samples n. The support of $\boldsymbol{v}_*$ is considered as successfully recovered if $\mathsf{supp}(\boldsymbol{v}_T) = T_*$. The success probability is then computed as the ratio of the number of trials that successfully recover the support over 50 independent trials. For a fixed small sample size, we observe that using tree-sparse projection achieves higher success probability of support recovery compared with using the vanilla k-sparse projection.

Algorithm 4 Projected Power Method for Rank-r Case

Input: Sample covariance matrix $\widehat{\mathbf{\Sigma}}$, rank parameter r.

1: Set $\widehat{\mathbf{\Sigma}}^{(1)} = \widehat{\mathbf{\Sigma}}$.
2: **for** $\ell = 1, \ldots, r$ **do**
3: Compute $\tilde{\boldsymbol{v}}^{(\ell)} \leftarrow$ Algorithm 1 with current input covariance $\widehat{\mathbf{\Sigma}}^{(\ell)}$.
4: Orthogonalized current point

$$\boldsymbol{v}^{(\ell)} \leftarrow \frac{(\boldsymbol{I}_d - \boldsymbol{Q}_{\ell-1}\boldsymbol{Q}_{\ell-1}^\top)\tilde{\boldsymbol{v}}^{(\ell)}}{\|(\boldsymbol{I}_d - \boldsymbol{Q}_{\ell-1}\boldsymbol{Q}_{\ell-1}^\top)\tilde{\boldsymbol{v}}^{(\ell)}\|_2}$$

with $\boldsymbol{Q}_{\ell-1} := (\boldsymbol{v}^{(1)}|\cdots|\boldsymbol{v}^{(\ell-1)})$ an orthogonal matrix.
5: Update covariance $\widehat{\mathbf{\Sigma}}^{(\ell+1)} \leftarrow \widehat{\mathbf{\Sigma}}^{(\ell)} - \lambda^{(\ell)}\boldsymbol{v}^{(\ell)}(\boldsymbol{v}^{(\ell)})^\top$ with $\lambda^{(\ell)} = (\boldsymbol{v}^{(\ell)})^\top\widehat{\mathbf{\Sigma}}^{(\ell)}\boldsymbol{v}^{(\ell)}$.
6: **end for**

Output: $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(r)}$.

forms of structured PCA [56]. Another natural direction is to consider more than a single principal component. Progress has been made towards establishing general fundamental limits in these settings [13]; there are also natural analogs for the projected power method in these settings and it would be interesting to analyze it under a general structural assumption along with statistical and computational limits. In particular, for finding top-rk-sparse principal components, i.e., rank-r case, similar to the spiked Wishart model used for rank-1 case, we assume the covariance matrix $\mathbf{\Sigma} = \sum_{j=1}^r \lambda_j \boldsymbol{v}_*^j(\boldsymbol{v}_*^j)^\top + \boldsymbol{I}$, where $\boldsymbol{v}_*^j \in \mathcal{M}$ for all $j \in [r]$, $\langle \boldsymbol{v}_*^j, \boldsymbol{v}_*^{j'} \rangle = 0$ with $j \neq j' \in [r]$ and $\lambda_1 > \cdots > \lambda_r > 0$ $(=: \lambda_{r+1})$ with a fixed positive eigengap $\Delta := \min_{j=1}^r \{\lambda_j - \lambda_{j+1}\} > 0$.

- **Without shared structure.** Suppose the top-r principal components do not have the same structure (i.e., we have $\boldsymbol{v}_*^j \in L^{(j)} \in \mathcal{M},\ \forall\ j \in [r]$, and corresponding linear subspace $L^{(j)}$ may not equal to $L^{(j')}$ when $j \neq j' \in [r]$), we can apply the proposed projected power method with an additional deflation method (see Section 2.3.3 of Mackey [40]). See Algorithm 4 below. An interesting direction for future work is to extend our analysis techniques to handle Algorithm 4.
- **With shared structure.** Suppose the top-r principal components have the same structure, that is to say, there exists a linear subspace $L_* \in \mathcal{M}$ such that all top-r principal components satisfy $\boldsymbol{v}_*^1, \ldots, \boldsymbol{v}_*^r \in L_*$. In this case, solving the exact projection subproblem

$$\operatorname*{argmin}_{\boldsymbol{U} \in \mathbb{R}^{d\times r}} \|\boldsymbol{V} - \boldsymbol{U}\|_F^2 \ \text{ s.t. } \boldsymbol{U} \in \mathcal{M},\ \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_r$$

is challenging even for vanilla sparse PCA, and we are not aware of an efficient algorithm.
Having said that, an inexact projected power method ensures local convergence for vanilla sparse PCA (see Theorem 3.1 of Ma [39]). It is an interesting open question whether inexact projections suffice for path-/tree-sparse PCA.

Algorithm 5 Intersection Verification for $\mathcal{B}$ or $\mathbf{\Phi}$

Input. $\mathcal{L}$, each linear subspace $L \in \mathcal{L}$ is represented by $\dim(L)$ independent vectors in L.

1: **Initialize** $\mathcal{L}^{(0)} := \emptyset, \mathcal{B}^{(0)} := \emptyset, t = 0$.
2: **while** $|\mathcal{B}^{(t)}| < d$ **do** **outer while-loop**
3: Pick a linear subspace $L^{(t)} \in \mathcal{L}\backslash\mathcal{L}^{(t)}$.
4: **while** True **do** **inner while-loop**
5: Select $\tilde{L}^{(t)} \in \mathcal{L}\backslash\{L^{(t)}\}$ uniformly at random without replacement.
6: **if** $L^{(t)} \cap \tilde{L}^{(t)} \neq \{\mathbf{0}\}$ **then**
7: Compute three bases for $L^{(t)}, \tilde{L}^{(t)}$.
8: Update $\mathcal{B}^{(t+1)}$ via adding the above three bases.
9: Update $\mathcal{L}^{(t+1)} := \mathcal{L}^{(t)} \cup \{L^{(t)}, \tilde{L}^{(t)}\}$.
10: **Break** inner while-loop.
11: **end if**
12: **end while**
13: **end while**

Output. $\mathcal{B}^{(t+1)}$ or $\mathbf{\Phi}$ with columns all bases in $\mathcal{B}^{(t+1)}$.

APPENDIX

A. Time-Consuming Case in Section II

Example 1: **Time-Consuming Case.** Given $\mathcal{L} = \{L_1, \ldots, L_{d-1}, L_d\}$ with $L_i = \operatorname{span}(\phi_i, \phi_{i+1})$ for $i = 1, \ldots, d-1$ and $L_d = \operatorname{span}(\phi_1, \phi_d)$. Each linear subspace L_i is known by given two linearly independent but not necessarily orthonormal vectors, say $\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}$, in L_i. As a result, for a given linear subspace L, we do not know the index $i \in [d]$ of this linear subspace L based on the given vectors $\boldsymbol{u}_1^{(\cdot)}, \boldsymbol{u}_2^{(\cdot)} \in L$. Hence the corresponding two bases that spans this known linear subspace L is unknown to us.

From the Example 1, if two linear subspaces L, L' have a non-zero intersection, i.e., $L \cap L' \neq \{\mathbf{0}\}$, then the base $\phi = L \cap L' \in \mathcal{B}$ is uniquely determined, and so as the rest two bases in L, L' respectively. Thus computing $\mathbf{\Phi}$ from $\mathcal{L} := \{L_1, \ldots, L_{d-1}, L_d\}$ is equivalent to find out all bases $\phi \in \mathcal{B}$ via intersection verification. Since we do not know the index corresponding to each linear subspace, to compute one base in $\mathcal{B}$, what we can do is to verify the intersection of two randomly chosen linear subspaces. The detailed procedures of computing the unknown orthonormal basis $\mathcal{B}$ are presented in the randomized algorithm 5.

Proposition 3: **Expected Running Time of Algorithm 5.** Under the setting of $\mathcal{L}$ presented in Example 1, the expected running time of Algorithm 5 is of order $O(d^3)$.

Thus, finding all bases takes more than $d^2/9$ intersection verifications in expectation. Each intersection verification requires $O(d)$ time. Then the expected running time of computing $\mathbf{\Phi}$ is $O(d^3)$. In contrast, computing the exact projection of $\boldsymbol{v}$ onto $\mathcal{M}$ takes $O(d^2)$ running time[5]. Therefore, the above analysis illustrates that extracting $\mathbf{\Phi}$ takes way more time than just

[5]Projecting onto a 1D linear subspace takes $O(d)$ time, and there are d linear subspaces in total.

implementing the projection, which further explains why $\mathbf{\Phi}$ is not necessary to recover the true PC $\boldsymbol{v}_*$.

Moreover, under the general setting of $\mathcal{L} = \{L_1, \ldots, L_M\}$, given a set of independent and not necessarily orthonormal vectors $\boldsymbol{u}_1^{(m)}, \ldots, \boldsymbol{u}_{\dim(L_m)}^{(m)}$ of each linear subspace L_m with $m \in [M]$, it is unclear whether and how long one could find the orthonormal basis $\mathbf{\Phi}$ from $\mathcal{L}$ via solving the following variant of dictionary learning problem (18),

$$\begin{aligned} &\min_{\mathbf{\Phi}, \boldsymbol{R}} \left\| [\boldsymbol{U}^{(1)} \mid \cdots \mid \boldsymbol{U}^{(M)}] - \mathbf{\Phi}[\boldsymbol{R}^{(1)} \mid \cdots \mid \boldsymbol{R}^{(M)}] \right\|_F^2, \\ &\text{s.t.} \quad \mathbf{\Phi}\mathbf{\Phi}^\top = \boldsymbol{I}_d, \quad \|\boldsymbol{R}^{(m)}\|_0 \le \dim(L_m) \ \ \forall\, m \in M \end{aligned} \tag{18}$$

where, for all $m \in [M]$, $\boldsymbol{U}^{(m)}$ denotes the matrix with columns $\boldsymbol{u}_1^{(m)}, \ldots, \boldsymbol{u}_{\dim(L_m)}^{(m)}$ and $\|\boldsymbol{R}^{(m)}\|_0 \le \dim(L_m)$ denotes that the number of non-zero rows of $\boldsymbol{R}^{(m)}$ is at most $\dim(L_m)$.

Proof of Proposition 3: First, based on the setting of each L_i for $i = 1, \ldots, d$, L_i has non-zero intersection with L_{i-1} and L_{i+1}. Thus the expected number of selections (i.e., inner while-loop (4)) for step (5) of Algorithm 5 satisfies

$$\begin{aligned} &\mathbb{E}[\text{number of selections}] \\ &= 1 \cdot \frac{2}{d-1} + 2 \cdot \frac{d-3}{d-1}\frac{2}{d-2} + 3 \cdot \frac{d-3}{d-1}\frac{d-4}{d-2}\frac{2}{d-3} + \cdots \\ &= \sum_{i=1}^{d-2} i \cdot \frac{2(d-i-1)}{(d-1)(d-2)} = \frac{d}{3}. \end{aligned}$$

Every time we find $L^{(t)} \cap \tilde{L}^{(t)} \neq \{\mathbf{0}\}$, in step (8) of Algorithm 5, we can add three more new bases to $\mathcal{B}^{(t+1)}$ if $\tilde{L}^{(t)} \notin \mathcal{L}^{(t)}$, and one more new base to $\mathcal{B}^{(t+1)}$ if $\tilde{L}^{(t)} \in \mathcal{L}^{(t)}$. Therefore, the number of outer while-loop (2) of Algorithm 5 satisfies

$$\begin{aligned} &\text{number of outer while-loop} \\ &= \text{selection with 3 more bases} + \text{selection with 1 more bases}. \end{aligned}$$

Moreover, the stopping criteria of outer while-loop (2) of Algorithm 5 ensures that the number of outer while-loop (2) is greater than or equal to $d/3$, where the equality holds when we can add three more new bases at every inner while-loop (4) of Algorithm 5. Therefore, in expectation, the total number of selections of Algorithm 5 satisfies

$$\begin{aligned} &\mathbb{E}[\text{total number of selections}] \\ &= \text{number of outer while-loop} \times \mathbb{E}[\text{number of selections}] \\ &\ge d^2/9. \end{aligned}$$

Since we do an intersection verification for each selection, and an intersection verification takes $O(d)$ running time, then the expected total running time of computing $\mathbf{\Phi}$ is $O(d^3)$. ■

ACKNOWLEDGMENT

We are grateful to the Simons Institute for the Theory of Computing for their hospitality, where part of this work was performed.

REFERENCES

[1] O. Alter, P. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc. Nat. Acad. Sci.*, vol. 97, no. 18, pp. 10101–10106, 2000.
[2] A. Amini and M. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," in *Proc. IEEE Int. Symp. Inf. Theory*, Piscataway, NJ, USA: IEEE Press, 2008, pp. 2454–2458.
[3] M. Asteris, A. Kyrillidis, A. Dimakis, H.-G. Yi, and B. Chandrasekaran, "Stay on path: PCA along graph paths," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2015, pp. 1728–1736.
[4] A. Bandeira, D. Kunisky, and A. Wein, "Computational hardness of certifying bounds on constrained pca problems," 2019, *arXiv:1902.07324*.
[5] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.
[6] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Proc. Conf. Learn. Theory*, PMLR, 2013, pp. 1046–1066.
[7] T. Bie, J. Suykens, and B. Moor, "Learning from general label constraints," in *Proc. Joint IAPR Int. Workshops Statist. Techn. Pattern Recognit. (SPR) Struct. Syntactic Pattern Recognit. (SSPR)*, New York, NY, USA: Springer-Verlag, 2004, pp. 671–679.
[8] A. Birnbaum, I. Johnstone, B. Nadler, and D. Paul, "Minimax bounds for sparse PCA with noisy high-dimensional data," *Ann. Statist.*, vol. 41, no. 3, p. 1055, 2013.
[9] M. Brennan and G. Bresler, "Optimal average-case reductions to sparse PCA: From weak assumptions to strong hardness," in *Proc. Conf. Learn. Theory*, PMLR, 2019, pp. 469–470.
[10] M. Brennan and G. Bresler, "Reducibility and statistical-computational gaps from secret leakage," in *Proc. Conf. Learn. Theory*, PMLR, 2020, pp. 648–847.
[11] M. Brennan, G. Bresler, and W. Huleihel, "Reducibility and computational lower bounds for problems with planted sparse structure," in *Proc. Conf. Learn. Theory*, PMLR, 2018, pp. 48–166.
[12] J. Cadima and I. Jolliffe, "Loading and correlations in the interpretation of principle compenents," *J. Appl. Statist.*, vol. 22, no. 2, pp. 203–214, 1995.
[13] T. Cai, H. Li, and R. Ma, "Optimal structured principal subspace estimation: Metric entropy and minimax rates," *J. Mach. Learn. Res.*, vol. 22, no. 46, pp. 1–45, 2021.
[14] T. Cai, Z. Ma, and Y. Wu, "Sparse PCA: Optimal rates and adaptive estimation," *Ann. Statist.*, vol. 41, no. 6, pp. 3074–3110, 2013.
[15] C. Cartis and A. Thompson, "An exact tree projection algorithm for wavelets," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1026–1029, Nov. 2013.
[16] S. Chen, S. Ma, L. Xue, and H. Zou, "An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis," *INFORMS J. Optim.*, vol. 2, no. 3, pp. 192–208, 2020.
[17] A. d'Aspremont, L. Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," in *Proc. Adv. Neural Inf. Process. Syst. 17*.
[18] Y. Deshpande and A. Montanari, "Sparse PCA via covariance thresholding," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 4913–4953, 2016.
[19] Y. Deshpande, A. Montanari, and E. Richard, "Cone-constrained principal component analysis," in *Proc. Adv. Neural Inf. Process. Syst. 27*, 2014.
[20] S. Dey, R. Mazumder, and G. Wang, "Using L1-relaxation and integer programming to obtain dual bounds for sparse PCA," Operations Research, vol. 70, no. 3, pp. 1914–1932, 2022.
[21] S. Dey, M. Molinaro, and G. Wang, "Solving sparse principal component analysis with global support," *Math. Program.*, vol. 199, no. 1, pp. 421–459, 2023.
[22] L. Ramesh, C. R. Murthy, and H. Tyagi, "Multiple support recovery using very few measurements per sample," *IEEE Trans. Sig. Process.*, vol. 70, pp. 2193–2206, 2022.
[23] A. d'Aspremont, F. Bach, and L. Ghaoui, "Approximation bounds for sparse principal component analysis," *Math. Program.*, vol. 148, no. 1, pp. 89–110, 2014.
[24] N. Erichson, P. Zheng, K. Manohar, S. Brunton, J. Kutz, and A. Aravkin, "Sparse principal component analysis via variable projection," *SIAM J. Appl. Math.*, vol. 80, no. 2, pp. 977–1002, 2020.
[25] C.-M. Feng, Y. Xu, J.-X. Liu, Y.-L. Gao, and C.-H. Zheng, "Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data," *IEEE Trans. neural networks Learn. Syst.*, vol. 30, no. 10, pp. 2926–2937, Oct. 2019.

[26] G. Frusque, J. Jung, P. Borgnat, and P. Gonçalves, "Sparse tensor dimensionality reduction with application to clustering of functional connectivity," in *Proc. Wavelets Sparsity XVIII*, vol. 11138. California, USA: International Society for Optics and Photonics, 2019, pp. 201–217.
[27] C. Gao, Z. Ma, and H. Zhou, "Sparse CCA: Adaptive estimation and computational barriers," *Ann. Statist.*, vol. 45, no. 5, pp. 2074–2101, 2017.
[28] P. Hancock, A. Burton, and V. Bruce, "Face processing: Human perception and principal components analysis," *Memory Cognit.*, vol. 24, no. 1, pp. 26–40, 1996.
[29] T. Hastie et al., "'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns," *Genome Biol.*, vol. 1, no. 2, pp. 1–21, 2000.
[30] T. Hastie, R. Tibshirani, J. Friedman, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. New York, NY, USA: Springer-Verlag, 2009.
[31] C. Hegde, P. Indyk, and L. Schmidt, "A nearly-linear time framework for graph-structured sparsity," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2015, pp. 928–937.
[32] I. Johnstone and A. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *J. Amer. Statist. Assoc.*, vol. 104, no. 486, pp. 682–693, 2009.
[33] I. Jolliffe, N. Trendafilov, and M. Uddin, "A modified principal component technique based on the LASSO," *J. Comput. Graphical Statist.*, vol. 12, no. 3, pp. 531–547, 2003.
[34] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learn. Res.*, vol. 11, no. 2, 2010.
[35] J. Kim, M. Tawarmalani, and J. Richard, "Convexification of permutation-invariant sets and applications," Scanning Electron Microsc Meet at 2019.
[36] Y. Li and W. Xie, "Exact and approximation algorithms for sparse PCA," 2020, *arXiv:2008.12438*.
[37] Z. Liu, J. Liu, S. Ghosh, J. Han, and J. Scarlett, "Generative principal component analysis," in *Proc. Int. Conf. Learn. Representations*, 2021.
[38] T. Ma and A. Wigderson, "Sum-of-squares lower bounds for sparse PCA," in *Proc. Adv. Neural Inf. Process. Syst. 28*, 2015.
[39] Z. Ma, "Sparse principal component analysis and iterative thresholding," *Ann. Statist.*, vol. 41, no. 2, pp. 772–801, 2013.
[40] L. Mackey, "Deflation methods for sparse PCA," in *Proc. Adv. Neural Inf. Process. Syst. 21*, Dec. 2008, pp. 1017–1024.
[41] S. Mallat, *A Wavelet Tour of Signal Processing*. Amsterdam, The Netherlands: Elsevier, 1999.
[42] A. Moitra, W. Perry, and A. Wein, "How robust are reconstruction thresholds for community detection?" in *Proc. 48th Annu. ACM Symp. Theory Comput.*, 2016, pp. 828–841.
[43] A. Montanari and E. Richard, "Non-negative principal component analysis: Message passing algorithms and sharp asymptotics," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1458–1484, Mar. 2016.
[44] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc.: Ser. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.
[45] L. Tran, L. Tran, T. Hoang, and B. Bui, "Tensor sparse PCA and face recognition: A novel approach," *SN Appl. Sci.*, vol. 2, no. 7, pp. 1–7, 2020.
[46] S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1247–1251, Jul. 1994.
[47] V. Vu and J. Lei, "Minimax rates of estimation for sparse PCA in high dimensions," in *Proc. Artif. Intell. Statist.*, PMLR, 2012, pp. 1278–1286.
[48] V. Vu, J. Cho, J. Lei, and K. Rohe, "Fantope projection and selection: A near-optimal convex relaxation of sparse PCA," in *Proc. Adv. Neural Inf. Process. Syst, 26*, 2013.
[49] V. Vu and J. Lei, "Minimax sparse principal subspace estimation in high dimensions," *Ann. Statist.*, vol. 41, no. 6, pp. 2905–2947, 2013.
[50] M. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[51] G. Wang, M. Lou, and A. Pananjady, "Do algorithms and barriers for sparse principal component analysis extend to other structured settings?" 2023, *arXiv:2307.13535*.
[52] T. Wang, Q. Berthet, and R. Samworth, "Statistical and computational trade-offs in estimation of sparse principal components," *Ann. Statist.*, vol. 44, no. 5, pp. 1896–1930, 2016.
[53] Z. Wang, B. Liu, S. Chen, S. Ma, L. Xue, and H. Zhao, "A manifold proximal linear method for sparse spectral clustering with application to single-cell RNA sequencing data analysis," *INFORMS J. Optim.*, 2021, *arXiv:2007.09524*.
[54] D. 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.
[55] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," in *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.
[56] Y. Yi and M. Neykov, "Non-sparse PCA in high dimensions via cone projected power iteration," 2020, *arXiv:2005.07587*.
[57] B. Yu, "Assouad, fano, and le cam," in *Festschrift for Lucien Le Cam*. New York, NY, USA: Springer-Verlag, 1997, pp. 423–435.
[58] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.
[59] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *J. Mach. Learn. Res.*, vol. 14, no. 28, pp. 899–925, 2013.
[60] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Adv. Phys.*, vol. 65, no. 5, pp. 453–552, 2016.
[61] Y. Zhang, A. d'Aspremont, and L. Ghaoui, "Sparse PCA: Convex relaxations, algorithms and applications," in *Handbook on Semidefinite, Conic and Polynomial Optimization*. New York, NY, USA: Springer-Verlag, 2012, pp. 915–940.
[62] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B (Statist. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.
[63] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graphical Statist.*, vol. 15, no. 2, pp. 265–286, 2006.

Guanyi Wang received the bachelor's degree in mathematics from the University of Science and Technology Beijing, the master's degree in applied mathematics and statistics from the Department of Applied Mathematics and Statistics, Johns Hopkins University, advised by Dr. Amitabh Basu, and the Ph.D. degree in algorithms, combinatorics and optimization (ACO) from Milton Stewart School of Industrial and Systems Engineering (ISyE), Georgia Institute of Technology, in 2021, advised by Dr. Santanu S. Dey. He is an Assistant Professor with the Department of Industrial Systems Engineering and Management, National University of Singapore (NUS). His research interests include in the area of mixed-integer nonlinear programming (MINLP) and statistical learning. His research is partly motivated by applications of MINLP arising in areas such as transportation, fairness in decision-making, statistics, and machine learning. Before joining the NUS, he was a Postdoctoral Researcher with the Polytechnique Montréal, under the supervision of Dr. Andrea Lodi from 2021 to 2022.

Mengqi Lou received the bachelor's degree in the mechanical engineering from Zhejiang University, in 2017, and the master's degree in robotics from Johns Hopkins University, in 2020. He is currently working toward the Ph.D. degree in algorithms, combinatorics and optimization with Georgia Institute of Technology, affiliated with the School of Industrial and Systems Engineering. His research interests include in the intersection of statistics, optimization, and applied probability.

Ashwin Pananjady received the B.Tech. degree in electrical engineering from IIT Madras, in 2014, and the Ph.D. degree in electrical engineering and computer sciences (EECS) from the University of California, Berkeley, in 2020. He is currently an Assistant Professor with Georgia Institute of Technology with a joint appointment between the School of Industrial and Systems Engineering and the School of Electrical and Computer Engineering. His research interests include high-dimensional statistics, information theory, and optimization. He has won the Inaugural Lawrence D. Brown Ph.D. Student Award from the Institute of Mathematical Statistics, a Young Researchers Prize (runner-up) from the Mathematical Optimization Society, the David J. Sakrison Memorial Prize (EECS Dissertation Award, UC Berkeley), and the Simons-Berkeley Research Fellowship in probability, geometry and computation in high dimensions.