Nonnegative Sparse PCA with Provable Guarantees

Megasthenis Asteris MEGAS@UTEXAS.EDU
Dimitris S. Papailiopoulos DIMITRIS@UTEXAS.EDU
Alexandros G. Dimakis DIMAKIS@AUSTIN.UTEXAS.EDU
Department of Electrical and Computer Engineering, The University of Texas at Austin, TX, USA

Abstract

We introduce a novel algorithm to compute nonnegative sparse principal components of positive semidefinite (PSD) matrices. Our algorithm comes with approximation guarantees contingent on the spectral profile of the input matrix $\mathbf{A}$: the sharper the eigenvalue decay, the better the quality of the approximation.

If the eigenvalues decay like any asymptotically vanishing function, we can approximate nonnegative sparse PCA within any accuracy ϵ in time polynomial in the matrix dimension n and desired sparsity k, but not in $1/\epsilon$. Further, we obtain a data dependent bound that is computed by executing an algorithm on a given data set. This bound is significantly tighter than *a-priori* bounds and can be used to show that for all tested datasets our algorithm is provably within $40\% - 90\%$ from the unknown optimum.

Our algorithm is combinatorial and explores a subspace defined by the leading eigenvectors of $\mathbf{A}$. We test our scheme on several data sets, showing that it matches or outperforms the previous state of the art.

1. Introduction

Given a data matrix $\mathbf{S} \in \mathbb{R}^{n\times m}$ comprising m zero-mean vectors on n features, the first principal component (PC) is

$$\arg \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T \mathbf{A} \mathbf{x}, \tag{1}$$

where $\mathbf{A} = 1/m \cdot \mathbf{S}\mathbf{S}^T$ is the $n \times n$ positive semidefinite (PSD) empirical covariance matrix. Subsequent PCs can be computed after $\mathbf{A}$ has been appropriately deflated to remove the first eigenvector. PCA is arguably the workhorse of high dimensional data analysis and achieves dimensionality reduction by computing the directions of maximum variance. Typically, all n features affect positively or negatively these directions resulting in dense PCs, which explain the largest possible data variance, but are often not interpretable.

It has been shown that enforcing nonnegativity on the computed principal components can aid interpretability. This is particularly true in applications where features interact only in an additive manner. For instance, in bioinformatics, chemical concentrations are nonnegative (Kim & Park, 2007), or the expression level of genes is typically attributed to positive or negative influences of those genes, but not both (Badea & Tilivea, 2005). Here, enforcing nonnegativity, in conjunction with sparsity on the computed components can assist the discovery of local patterns in the data. In computer vision, where features may coincide with non negatively valued image pixels, nonnegative sparse PCA pertains to the extraction of the most informative image parts (Lee & Seung, 1999). In other applications, nonnegative weights admit a meaningful probabilistic interpretation.

Sparsity emerges as an additional desirable trait of the computed components because it further helps interpretability (Zou et al., 2006; d'Aspremont et al., 2007b), even independently of nonnegativity. From a machine learning perspective, enforcing sparsity serves as an unsupervised feature selection method: the active coordinates in an optimal l_0-norm constrained PC should correspond to the most informative subset of features. Although nonnegativity inherently promotes sparsity, an explicit sparsity constraint enables precise control on the number of selected features.

Nonnegative Sparse PC. Nonnegativity and sparsity can be directly enforced on the principal component optimization by adding constraints to (1). The k-sparse nonnegative principal component of $\mathbf{A}$ is

$$\mathbf{x}_\star = \arg \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T \mathbf{A} \mathbf{x}, \tag{2}$$

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

where $\mathbb{S}_k^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1, \|\mathbf{x}\|_0 \leq k, \mathbf{x} \geq 0\}$, for a desired sparsity parameter $k \in [n]$.

The problem of computing the first eigenvector (1) is easily solvable, but with the additional sparsity and nonnegativity constraints problem (2) becomes computationally intractable. The cardinality constraint alone renders sparse PCA NP-hard (Moghaddam et al., 2006b). Even if the l_0-norm constraint is dropped, we show that problem (2) remains computationally intractable by reducing it to checking matrix copositivity, a well known co-NP complete decision problem (Murty & Kabadi, 1987; Parrilo, 2000). Therefore, each of the constraints $\mathbf{x} \geq \mathbf{0}$ and $\|\mathbf{x}\|_0 \leq k$ individually makes the problem intractable.

Our Contribution: We introduce a novel algorithm for approximating the nonnegative k-sparse principal component with provable approximation guarantees.

Given any PSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, sparsity parameter k, and accuracy parameter $d \in [n]$, our algorithm outputs a nonnegative, k-sparse, unit norm vector $\mathbf{x}_d$ that achieves at least ρ_d fraction of the maximum objective value in (2), *i.e.*,

$$\mathbf{x}_d^T \mathbf{A} \mathbf{x}_d \;\geq\; \rho_d \cdot \mathbf{x}_\star{}^T \mathbf{A} \mathbf{x}_\star, \tag{3}$$

where

$$\rho_d \;\geq\; \max\left\{ \frac{k}{2n},\; \frac{1}{1 + 2\frac{n}{k}\lambda_{d+1}/\lambda_1} \right\}. \tag{4}$$

Here, λ_i is the i^{th} largest eigenvalue of $\mathbf{A}$, and the accuracy parameter d specifies the rank of the approximation used and controls the running time. Specifically, our algorithm runs in time $O(n^d k^d + n^{d+1})$. As can be seen our result depends on the spectral profile of $\mathbf{A}$: the faster the eigenvalue decay, the tighter the approximation.

Near-Linear time approximation. Our algorithm has a running time $O(n^d k^d + n^{d+1})$, which in the linear sparsity regime can be as high as $O(n^{2d})$. This can be non-practical for large data sets, even if we set the rank parameter d to be two or three. We present a modification of our algorithm that can provably approximate the result of the first in near-linear time. Specifically, for any desired accuracy $\epsilon \in (0, 1]$ it computes a nonnegative, k-sparse, unit norm vector $\widehat{\mathbf{x}}_d$ such that

$$\widehat{\mathbf{x}}_d^T \mathbf{A} \widehat{\mathbf{x}}_d \;\geq\; (1 - \epsilon) \cdot \rho_d \cdot \mathbf{x}_\star{}^T \mathbf{A} \mathbf{x}_\star, \tag{5}$$

where ρ_d is as described in (4). We show that the running time of our approximate algorithm is $O\left(\epsilon^{-d} \cdot n \log n\right)$, which is near-linear in n for any fixed accuracy parameters d and ϵ.

Our approximation theorem has several implications.

Exact solution for low-rank matrices. Observe that if the matrix $\mathbf{A}$ has rank d, our algorithm returns the optimal k-sparse PC for any target sparsity k. The same holds in the case of the rank-d update matrix $\mathbf{A} = \sigma\mathbf{I} + \mathbf{C}$, with $\text{rank}(\mathbf{C}) = d$ and arbitrary constant σ, since the algorithm can be equivalently applied on $\mathbf{C}$.

PTAS for any spectral decay. Consider the linear sparsity regime $k = c \cdot n$ and assume that the eigenvalues follow a decay law $\lambda_i \leq \lambda_1 \cdot f(i)$ for any decay function $f(i)$ which vanishes: $f(i) \to 0$ as $i \to \infty$. Special cases include power law decay $f(i) = 1/i^\alpha$ or even very slow decay functions like $f(i) = 1/\log\log i$. For all these cases, we can solve nonnegative sparse PCA for any desired accuracy ϵ in time polynomial in n and k, but not in $1/\epsilon$. Therefore, we obtain a polynomial-time approximation scheme (PTAS) for any spectral decay behavior.

Computable upper bounds. In addition to these theoretical guarantees, our method yields a data dependent upper bound on the maximum value of (2), that can be computed by running our algorithm. As it can be seen in Fig. 4-6, the obtained upper bound, combined with our achievable point, sandwiches the unknown optimum within a narrow region. Using this upper bound we are able to show that our solutions are within $40 - 90\%$ from the optimal in all the datasets that we examine. To the best of our knowledge, this framework of data dependent bounds has not been considered in the previous literature.

1.1. Related Work

There is a substantial volume of work on sparse PCA, spanning a rich variety of approaches: from early heuristics in (Jolliffe, 1995), to the LASSO based techniques in (Jolliffe et al., 2003), the elastic net l_1-regression in (Zou et al., 2006), a greedy branch-and-bound technique in (Moghaddam et al., 2006a), or semidefinite programming approaches (d'Aspremont et al., 2008; Zhang et al., 2012; d'Aspremont et al., 2007a). This line of work does not consider or enforce nonnegativity constraints.

When nonnegative components are desired, fundamentally different approaches have been used. Nonnegative matrix factorization (Lee & Seung, 1999) and its sparse variants (Hoyer, 2004; Kim & Park, 2007) fall within that scope: data is expressed as (sparse) nonnegative linear combinations of (sparse) nonnegative parts. These approaches are interested in finding a lower dimensionality representation of the data that reveals latent structure and minimizes a reconstruction error, but are not explicitly concerned with the statistical significance of individual output vectors.

Nonnegativity as an additional constraint on (sparse) PCA first appeared in (Zass & Shashua, 2007). The authors suggested a coordinate-descent scheme that jointly computes a set of nonnegative sparse principal components, maximizing the cumulative explained variance. An l_1-penalty promotes sparsity of computed components on average,

but not on each component individually. A second convex penalty is incorporated to favor orthogonal components.

Similar convex optimization approaches for nonnegative PCA have been subsequently proposed in the literature. In (Allen & Maletić-Savatić, 2011) for instance, the authors suggest an alternating maximization scheme for the computation of the first nonnegative PC, allowing the incorporation of known structural dependencies.

A competitive algorithm for nonnegative sparse PCA was established in (Sigg & Buhmann, 2008), with the development of a framework stemming from Expectation-Maximization (EM) for a probabilistic generative model of PCA. The proposed algorithm, which enforces hard sparsity, or nonnegativity, or both constraints simultaneously, computes the first approximate PC in $O(n^2)$, *i.e.*, time quadratic in the number of features.

To the best of our knowledge, no prior works provide provable approximation guarantees for the nonnegative sparse PCA optimization problem. Further, no data dependent upper bounds have been present in the previous literature.

Differences from SPCA work. Our work is closely related to (Karystinos & Liavas, 2010; Asteris et al., 2011; Papailiopoulos et al., 2013) that introduced the ideas of solving low-rank quadratic combinatorial optimization problems on low-rank PSD matrices using hyperspectral transformations. Such transformations are called spannograms and follow a similar architecture. In this paper, we extend the spannogram framework to nonnegative sparse PCA. The most important technical issue compared to (Asteris et al., 2011; Papailiopoulos et al., 2013) is introducing nonnegativity constraints in spannogram algorithms.

To understand how this changes the problem, notice that in the original sparse PCA problem without nonnegativity constraints, if the support is known, the optimal principal component supported on that set can be easily found. However, under nonnegativity constraints, the problem is hard even if the optimal support is known. This is the fundamental technical problem that we address in this paper. We show that if the involved subspace is low-dimensional, it is possible to solve this problem.

2. Algorithm Overview

Given an $n \times n$ PSD matrix $\mathbf{A}$, the desired sparsity k, and an accuracy parameter $d \in [n]$, our algorithm computes a *nonnegative, k-sparse, unit norm* vector $\mathbf{x}_d$ approximating the nonnegative, k-sparse PC of $\mathbf{A}$. We begin with a high-level description of the main steps of the algorithm.

Step 1. *Compute $\mathbf{A}_d$, the rank-d approximation of $\mathbf{A}$.* We compute $\mathbf{A}_d$, the best rank-d approximation of $\mathbf{A}$, zeroing out the $n-d$ trailing eigenvalues of $\mathbf{A}$, that is,

$$\mathbf{A}_d = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^T,$$

where λ_i is the i^{th} largest eigenvalue of $\mathbf{A}$ and $\mathbf{u}_i$ the corresponding eigenvector.

Algorithm 1 Spannogram Nonnegative Sparse PCA

input $\mathbf{A}$ ($n \times n$ PSD matrix), $k \in [n]$, $d \in [n]$.
1: $\mathbf{U}, \mathbf{\Lambda} \leftarrow \mathtt{svd}(\mathbf{A}, d)$
2: $\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$ $\{\mathbf{A}_d = \mathbf{V}\mathbf{V}^T\}$
3: $\mathcal{S}_d \leftarrow$ Spannogram($\mathbf{V}$, k) {Algo. 2}
4: $\mathcal{X}_d \leftarrow \{\}$ $\{|\mathcal{S}_d| \leq O(n^d)\}$
5: **for all** $\mathcal{I} \in \mathcal{S}_d$ **do**
6: $\quad \mathbf{c}^{(\mathcal{I})} \leftarrow \arg\max_{\substack{\|\mathbf{c}\|_2=1 \\ \mathbf{V}_{\mathcal{I}}\mathbf{c} \geq \mathbf{0}}} \| (\mathbf{V}_{\mathcal{I}}\mathbf{c}) \|_2^2$ {Sec. 5}
7: $\quad \mathbf{x}_{\mathcal{I}}^{(\mathcal{I})} \leftarrow |\mathbf{V}_{\mathcal{I}}\mathbf{c}|/\|\mathbf{V}_{\mathcal{I}}\mathbf{c}\|, \quad \mathbf{x}_{\mathcal{I}^c}^{(\mathcal{I})} \leftarrow \mathbf{0}$
8: $\quad \mathcal{X}_d \leftarrow \mathcal{X}_d \cup \{\mathbf{x}^{(\mathcal{I})}\}$
9: **end for** $\{|\mathcal{X}_d| \leq |\mathcal{S}_d|\}$
output $\mathbf{x}_d \leftarrow \arg\max_{\mathbf{x} \in \mathcal{X}_d} \mathbf{x}^T \mathbf{A}_d \mathbf{x}$

Step 2. *Compute $\mathcal{S}_d$, a set of $O(n^d)$ candidate supports.* Enumerating the $\binom{n}{k}$ possible supports for k-sparse vectors in $\mathbb{R}^n$ is computationally intractable. Using our *Spannogram* technique described in Section 4, we efficiently determine a collection $\mathcal{S}_d$ of support sets, with cardinality $|\mathcal{S}_d| \leq 2^d \binom{n+1}{d}$, that provably contains the support of the nonnegative, k-sparse PC of $\mathbf{A}_d$.

Step 3. *Compute $\mathcal{X}_d$, a set of candidate solutions.* For each candidate support set $\mathcal{I} \in \mathcal{S}_d$, we compute a candidate solution $\mathbf{x}$ supported only in $\mathcal{I}$:

$$\arg \max_{\substack{\|\mathbf{x}\|_2=1, \mathbf{x} \geq \mathbf{0}, \\ \text{supp}(\mathbf{x}) \subseteq \mathcal{I}}} \mathbf{x}^T \mathbf{A}_d \mathbf{x}. \tag{6}$$

The constant rank of $\mathbf{A}_d$ is essential in solving (6): the constrained quadratic maximization is in general NP-hard, even for a given support.

Step 4. *Output the best candidate solution in $\mathcal{X}_d$, i.e.*, the candidate that maximizes the quadratic form.

If multiple components are desired, the procedure is repeated after an appropriate deflation has been applied on $\mathbf{A}_d$ (Mackey, 2008). The steps are formally presented in Algorithm 1. A detailed description is the subject of subsequent sections.

2.1. Approximation Guarantees

Instead of the nonnegative, k-sparse, principal component $\mathbf{x}_\star$ of $\mathbf{A}$, which attains the optimal value $\mathsf{OPT} = \mathbf{x}_\star{}^T \mathbf{A} \mathbf{x}_\star$, our algorithm outputs a nonnegative, k-sparse, unit norm vector $\mathbf{x}_d$. We measure the quality of $\mathbf{x}_d$ as a surrogate of $\mathbf{x}_\star$ by the approximation factor $\mathbf{x}_d{}^T \mathbf{A} \mathbf{x}_d / \mathsf{OPT}$. Clearly,

the approximation factor takes values in $(0, 1]$, with higher values implying tighter approximation.

Theorem 1. *For any $n \times n$ PSD matrix $\mathbf{A}$, sparsity parameter k, and accuracy parameter $d \in [n]$, Alg. 1 outputs a nonnegative, k-sparse, unit norm vector $\mathbf{x}_d$ such that*

$$\mathbf{x}_d{}^T \mathbf{A} \mathbf{x}_d \;\geq\; \rho_d \cdot \mathbf{x}_\star{}^T \mathbf{A} \mathbf{x}_\star,$$

where

$$\rho_d \;\geq\; \max\left\{ \frac{k}{2n},\; \frac{1}{1 + 2\frac{n}{k}\lambda_{d+1}/\lambda_1} \right\},$$

in time $O(n^{d+1} + n^d k^d)$.

The approximation guarantee of Theorem 1 relies on establishing connections among the eigenvalues of $\mathbf{A}$, and the quadratic forms $\mathbf{x}_d{}^T \mathbf{A} \mathbf{x}_d$ and $\mathbf{x}_d{}^T \mathbf{A}_d \mathbf{x}_d$. The proof can be found in the supplemental material. The complexity of Algorithm 1 follows upon its detailed description.

3. Proposed Scheme

Our algorithm approximates the nonnegative, k-sparse PC of a PSD matrix $\mathbf{A}$ by computing the corresponding PC of $\mathbf{A}_d$, a rank-d surrogate of the input argument $\mathbf{A}$:

$$\mathbf{A}_d = \sum_{i=1}^{d} \mathbf{v}_i \mathbf{v}_i{}^T = \mathbf{V}\mathbf{V}^T, \tag{7}$$

where $\mathbf{v}_i = \sqrt{\lambda_i}\mathbf{u}_i$ is the scaled eigenvector corresponding to the i^{th} largest eigenvalue of $\mathbf{A}$, and $\mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_d] \in \mathbb{R}^{n \times d}$. In this section, we delve into the details of our algorithmic developments and describe how the low rank of $\mathbf{A}_d$ unlocks the computation of the desired PC.

3.1. Rank-1: A simple case

We begin with the rank-1 case because, besides its motivational simplicity, it is a fundamental component of the algorithmic developments for the rank-d case.

In the rank-1 case, $\mathbf{V}$ reduces to a single vector in $\mathbb{R}^n$ and $\mathbf{x}_1$, the nonnegative k-sparse PC of $\mathbf{A}_1$, is the solution to

$$\max_{\mathbf{x} \in \mathbb{S}_k^n} \mathbf{x}^T \mathbf{A}_1 \mathbf{x} = \max_{\mathbf{x} \in \mathbb{S}_k^n} \left(\mathbf{v}^T \mathbf{x}\right)^2. \tag{8}$$

That is, $\mathbf{x}_1$ is the nonnegative, k-sparse, unit length vector that maximizes $(\mathbf{v}^T\mathbf{x})^2$. Let $\mathcal{I} = \text{supp}(\mathbf{x}_1)$, $|\mathcal{I}| \leq k$, be the unknown support of $\mathbf{x}_1$. Then, $(\mathbf{v}^T\mathbf{x})^2 = \left(\sum_{i \in \mathcal{I}} v_i \cdot x_i\right)^2$. Since $\mathbf{x}_1 \geq \mathbf{0}$, it should not be hard to see that the active entries of $\mathbf{x}_1$ must correspond to nonnegative or nonpositive entries of $\mathbf{v}$, but not a combination of both. In other words, $\mathbf{v}_\mathcal{I}$, the entries of $\mathbf{v}$ indexed by $\mathcal{I}$, must satisfy $\mathbf{v}_\mathcal{I} \geq \mathbf{0}$ or $\mathbf{v}_\mathcal{I} \leq \mathbf{0}$. In either case, by the Cauchy-Schwarz inequality,

$$\left(\mathbf{v}^T\mathbf{x}\right)^2 = \left(\mathbf{v}_\mathcal{I}{}^T \mathbf{x}_\mathcal{I}\right)^2 \leq \|\mathbf{v}_\mathcal{I}\|_2^2 \|\mathbf{x}_\mathcal{I}\|_2^2, = \|\mathbf{v}_\mathcal{I}\|_2^2. \tag{9}$$

Equality in (9) can always be achieved by setting $\mathbf{x}_\mathcal{I} = \mathbf{v}_\mathcal{I}/\|\mathbf{v}_\mathcal{I}\|_2$ if $\mathbf{v}_\mathcal{I} \geq \mathbf{0}$, and $\mathbf{x}_\mathcal{I} = -\mathbf{v}_\mathcal{I}/\|\mathbf{v}_\mathcal{I}\|_2$ if $\mathbf{v}_\mathcal{I} \leq \mathbf{0}$. The support of the optimal solution $\mathbf{x}_1$ is the set $\mathcal{I}$ for which $\|\mathbf{v}_\mathcal{I}\|_2^2$ in (9) is maximized under the restriction that the entries of $\mathbf{v}_\mathcal{I}$ do not have mixed signs.

Def. 1. *Let $\mathcal{I}_k^+(\mathbf{v})$, $1 \leq k \leq n$ denote the set of indices of the (at most) k largest nonnegative entries in $\mathbf{v} \in \mathbb{R}^n$.*

Proposition 3.1. *Let $\mathbf{x}_1$ be the solution to problem (8). Then, supp $(\mathbf{x}_1) \in \mathcal{S}_1 = \left\{\mathcal{I}_k^+(\mathbf{v}),\, \mathcal{I}_k^+(-\mathbf{v})\right\}$.*

The collection $\mathcal{S}_1$ and the associated candidate vectors via (9) are constructed in $O(n)$. The solution $\mathbf{x}_1$ is the candidate that maximizes the quadratic.

3.2. Rank-d case

In the rank-d case, $\mathbf{x}_d$, the nonnegative, k-sparse PC of $\mathbf{A}_d$ is the solution to the following problem:

$$\max_{\mathbf{x} \in \mathbb{S}_k^n} \mathbf{x}^T \mathbf{A}_d \mathbf{x} = \max_{\mathbf{x} \in \mathbb{S}_k^n} \|\mathbf{V}^T\mathbf{x}\|_2^2. \tag{10}$$

Consider an auxiliary vector $\mathbf{c} \in \mathbb{R}^d$, with $\|\mathbf{c}\|_2 = 1$. From the Cauchy-Schwarz inequality,

$$\|\mathbf{V}^T\mathbf{x}\|_2^2 = \|\mathbf{c}\|_2^2 \|\mathbf{V}^T\mathbf{x}\|_2^2 \geq \left|\mathbf{c}^T \left(\mathbf{V}^T\mathbf{x}\right)\right|^2. \tag{11}$$

Equality in (11) is achieved if and only if $\mathbf{c}$ is colinear to $\mathbf{V}^T\mathbf{x}$. Since $\mathbf{c}$ spans the entire unit sphere, such a $\mathbf{c}$ exists for every $\mathbf{x}$, yielding an alternative description for the objective function in (10):

$$\|\mathbf{V}^T\mathbf{x}\|_2^2 = \max_{\mathbf{c} \in \mathbb{S}^d} \left|\left(\mathbf{V}\mathbf{c}\right)^T \mathbf{x}\right|^2, \tag{12}$$

where $\mathbb{S}^d = \left\{\mathbf{c} \in \mathbb{R}^d : \|\mathbf{c}\|_2 = 1\right\}$ is the d-dimensional unit sphere. The maximization in (10) becomes

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{S}_k^n} \|\mathbf{V}^T\mathbf{x}\|_2^2 &= \max_{\mathbf{x} \in \mathbb{S}_k^n} \max_{\mathbf{c} \in \mathbb{S}^d,} |\left(\mathbf{V}\mathbf{c}\right)^T \mathbf{x}|^2 \\ &= \max_{\mathbf{c} \in \mathbb{S}^d,\, \mathbf{x} \in \mathbb{S}_k^n} |\left(\mathbf{V}\mathbf{c}\right)^T \mathbf{x}|^2. \end{aligned} \tag{13}$$

The set of candidate supports. A first key observation is that for fixed $\mathbf{c}$, the product $(\mathbf{V}\mathbf{c})$ is a vector in $\mathbb{R}^n$. Maximizing $|\left(\mathbf{V}\mathbf{c}\right)^T \mathbf{x}|^2$ over all vectors $\mathbf{x} \in \mathbb{S}_k^n$ is a rank-1 instance of the optimization problem, as in (8). Let $(\mathbf{c}_d, \mathbf{x}_d)$ be the optimal solution of (10). By Proposition 3.1, the support of $\mathbf{x}_d$ coincides with either $\mathcal{I}_k^+(\mathbf{V}\mathbf{c}_d)$ or $\mathcal{I}_k^+(-\mathbf{V}\mathbf{c}_d)$. Hence, we can safely claim that supp$(\mathbf{x}_d)$ appears in

$$\mathcal{S}_d = \bigcup_{\mathbf{c} \in \mathbb{S}^d} \left\{\mathcal{I}_k^+(\mathbf{V}\mathbf{c})\right\}. \tag{14}$$

Naively, one might think that $\mathcal{S}_d$ can contain as many as $\binom{n}{k}$ distinct support sets. In Section 4, we show that $|\mathcal{S}_d| \leq 2^d\binom{n+1}{d}$ and present our Spannogram technique (Alg. 2) for efficiently constructing $\mathcal{S}_d$ in $O(n^{d+1})$. Each support in $\mathcal{S}_d$ corresponds to a candidate principal component.

Solving for a given support. We seek a pair $(\mathbf{x}, \mathbf{c})$ that maximizes (13) under the additional constraint that $\mathbf{x}$ is supported only on a given set $\mathcal{I}$. By the Cauchy-Schwarz inequality, the objective in (13) satisfies

$$|\left(\mathbf{V}\mathbf{c}\right)^T \mathbf{x}|^2 = |\left(\mathbf{V}_{\mathcal{I}}\mathbf{c}\right)^T \mathbf{x}_{\mathcal{I}}|^2 \leq \|\left(\mathbf{V}_{\mathcal{I}}\mathbf{c}\right)\|_2^2, \qquad (15)$$

where $\mathbf{V}_{\mathcal{I}}$ is the matrix formed by the rows of $\mathbf{V}$ indexed by $\mathcal{I}$. Equality in (15) is achieved if and only if $\mathbf{x}_{\mathcal{I}}$ is colinear to $\mathbf{V}_{\mathcal{I}}\mathbf{c}$. However, it is not achievable for arbitrary $\mathbf{c}$, as $\mathbf{x}_{\mathcal{I}}$ must be nonnegative. From Proposition 3.1, we infer that $\mathbf{x}$ being supported in $\mathcal{I}$ implies that all entries of $\mathbf{V}_{\mathcal{I}}\mathbf{c}$ have the same sign. Further, whenever the last condition holds, a nonnegative $\mathbf{x}_{\mathcal{I}}$ colinear to $\mathbf{V}_{\mathcal{I}}\mathbf{c}$ exists and equality in (15) can be achieved. Under the additional constraint that $\text{supp}(\mathbf{x}) = \mathcal{I} \in \mathcal{S}_d$, the maximization in (13) becomes

$$\max_{\mathbf{c}\in\mathbb{S}^d} \max_{\substack{\mathbf{x}\in\mathbb{S}_k^n \\ \text{supp}(\mathbf{x})\subseteq\mathcal{I}}} |\left(\mathbf{V}\mathbf{c}\right)^T \mathbf{x}|^2 = \max_{\substack{\mathbf{c}\in\mathbb{S}^d \\ \mathbf{V}_{\mathcal{I}}\mathbf{c}\geq\mathbf{0}}} \|\left(\mathbf{V}_{\mathcal{I}}\mathbf{c}\right)\|_2^2. \qquad (16)$$

The constraint $\mathbf{V}_{\mathcal{I}}\mathbf{c} \geq \mathbf{0}$ in (16), is equivalent to requiring that all entries in $\mathbf{V}_{\mathcal{I}}\mathbf{c}$ have the same sign, since $\mathbf{c}$ and $-\mathbf{c}$ achieve the same objective value.

The optimization problem in (16) is NP-hard. In fact, it encompasses the original nonnegative PCA problem as a special case. Here, however, the constant dimension $d = \Theta(1)$ of the unknown variable $\mathbf{c}$ permits otherwise intractable operations. In Section 5, we outline an $O(k^d)$ algorithm for solving this constrained quadratic maximization.

The algorithm. The previous discussion suggests a two-step algorithm for solving the rank-d optimization problem in (10). First, run the Spannogram algorithm to construct $\mathcal{S}_d$, the collection of $O(n^d)$ candidate supports for $\mathbf{x}_d$, in $O(n^{d+1})$. For each $\mathcal{I} \in \mathcal{S}_d$, solve (16) in $O(k^d)$ to obtain a candidate solution $\mathbf{x}^{(\mathcal{I})}$ supported on $\mathcal{I}$. Output the candidate solution that maximizes the quadratic $\mathbf{x}^T\mathbf{A}_d\mathbf{x}$. Efficiently combining the previous steps yields an $O(n^{d+1} + n^d k^d)$ procedure for approximating the nonnegative sparse PC, outlined in Alg. 1.

4. The Nonnegative Spannogram

In this section, we describe how to construct $\mathcal{S}_d$, the collection of candidate supports, defined in (14) as

$$\mathcal{S}_d = \bigcup_{\mathbf{c}\in\mathbb{S}^d} \left\{\mathcal{I}_k^+(\mathbf{V}\mathbf{c})\right\},$$

for a given $\mathbf{V} \in \mathbb{R}^{n\times d}$. $\mathcal{S}_d$ comprises all support sets induced by vectors in the range of $\mathbf{V}$. The *Spannogram* of $\mathbf{V}$ is a *visualization* of its range, and a valuable tool in efficiently collecting those supports.

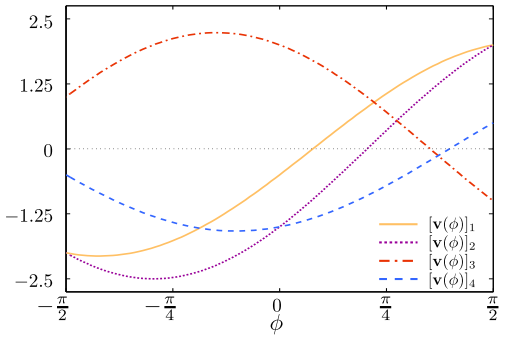


Figure 1. Spannogam of an arbitrary rank-2 matrix $\mathbf{V} \in \mathbb{R}^{4\times 2}$. At a point ϕ, the values of the curves correspond to the entries of a vector $\mathbf{v}(\phi)$ in the range of $\mathbf{V}$ and vice versa.

4.1. Constructing $\mathcal{S}_2$

We describe the $d = 2$ case, the simplest nontrivial case, to facilitate a gentle exposure to the Spannogram technique. The core ideas generalize to arbitrary d and a detailed description is provided in the supplemental material.

Spherical variables. Up to scaling, all vectors $\mathbf{v}$ in the range of $\mathbf{V} \in \mathbb{R}^{n\times 2}$, $\mathcal{R}(\mathbf{V})$, can be written as $\mathbf{v} = \mathbf{V}\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^2 : \|\mathbf{c}\| = 1$. We introduce a variable $\phi \in \Phi = (-\pi/2, \pi/2]$, and set $\mathbf{c}$ to be the following function of ϕ:

$$\mathbf{c}(\phi) = \left[\sin(\phi)\ \cos(\phi)\right]^T.$$

The range of $\mathbf{V}$, $\mathcal{R}(\mathbf{V}) = \{\pm\mathbf{v}(\phi) = \pm\mathbf{V}\mathbf{c}(\phi), \phi \in \Phi\}$, is also a function of ϕ, and in turn $\mathcal{S}_2$ can be expressed as

$$\mathcal{S}_2 = \bigcup_{\phi\in\Phi} \left\{\mathcal{I}_k^+(\mathbf{v}(\phi)),\ \mathcal{I}_k^+(-\mathbf{v}(\phi))\right\}.$$

Spannogram. The i^{th} entry of $\mathbf{v}(\phi)$ is a continuous function of ϕ generated by the i^{th} row of $\mathbf{V}$: $[\mathbf{v}(\phi)]_i = \mathbf{V}_{i,1}\sin(\phi) + \mathbf{V}_{i,2}\cos(\phi)$. Fig. 1 depicts the functions corresponding to the rows of an arbitrary matrix $\mathbf{V} \in \mathbb{R}^{4\times 2}$. We call this a *spannogram*, because at each ϕ, the values of the curves coincide with the entries of a vector in the range of $\mathbf{V}$. A key observation is that the sorting of the curves at some ϕ is locally invariant for most points in Φ. In fact, due to the continuity of the curves, as we move along the ϕ-axis, the set $\mathcal{I}_k^+(\mathbf{v}(\phi))$ can only change at points where a curve intersects with *(i)* another curve, or *(ii)* the zero axis; a change in either the sign of a curve or the relative order of two curves is necessary, although not sufficient, for $\mathcal{I}_k^+(\mathbf{v}(\phi))$ to change.

Appending a zero $(n+1)^{\text{th}}$ row to $\mathbf{V}$, the two aforementioned conditions can be merged into one: $\mathcal{I}_k^+(\mathbf{v}(\phi))$ can change only at the points where two of the $n+1$ curves intersect. Finding the unique intersection point of two curves $[\mathbf{v}(\phi)]_i$ and $[\mathbf{v}(\phi)]_j$ for all pairs $\{i, j\}$ is the key to dis-

covering all possible candidate support sets. There are exactly $\binom{n+1}{2}$ such points partitioning Φ into $\binom{n+1}{2}+1$ intervals within which the set of largest k nonnegative entries of $\mathbf{v}(\phi)$ and $-\mathbf{v}(\phi)$ are invariant.

Constructing $\mathcal{S}_2$. The point ϕ_{ij} where the i^{th} and j^{th} curves intersect, corresponds to a vector $\mathbf{v}(\phi_{ij}) \in \mathcal{R}(\mathbf{V})$ whose i^{th} and j^{th} entries are equal. To find it, it suffices to compute a $\mathbf{c} \neq \mathbf{0}$ such that $(\mathbf{e}_i - \mathbf{e}_j)^T\mathbf{V}\mathbf{c} = 0$, *i.e.*, a unit norm vector $\mathbf{c}_{ij}$ in the one-dimensional nullspace of $(\mathbf{e}_i - \mathbf{e}_j)^T\mathbf{V}$. Then, $\mathbf{v}(\phi_{ij}) = \mathbf{V}\mathbf{c}_{ij}$.

We compute the candidate support $\mathcal{I}_k^+(\mathbf{v}(\phi_{ij}))$ at the intersection. Assuming for simplicity that only the i^{th} and j^{th} curves intersect at ϕ_{ij}, the sorting of all curves is unchanged in a small neighborhood of ϕ_{ij}, except the i^{th} and j^{th} curves whose order changes over ϕ_{ij}. If both the i^{th} and j^{th} entries of $\mathbf{v}(\phi_{ij})$ or none of them is included in the k largest nonnegative entries, then the set $\mathcal{I}_k^+(\mathbf{v}(\phi))$ in the two intervals incident to ϕ_{ij} is identical. Otherwise, the i^{th} and j^{th} curve occupy the k^{th} and $(k+1)^{\text{th}}$ order at ϕ_{ij}, and the change in their relative order implies that one leaves and one joins the set of k largest nonnegative curves at ϕ_{ij}. The support sets associated with the two adjacent intervals differ only in one element (one contains index i and the other contains index j instead), while the remaining $k-1$ common indices correspond to the $k-1$ largest curves at the intersection point ϕ_{ij}. We include both in $\mathcal{S}_2$ and repeat the above procedure for $\mathcal{I}_k^+(-\mathbf{v}(\phi_{ij}))$.

Each pairwise intersection is computed in $O(1)$ and the at most 4 associated candidate supports in $O(n)$. In total, the collection $\mathcal{S}_2$ comprises $|\mathcal{S}_2| \leq 4\binom{n+1}{2} = O(n^2)$ candidate supports and can be constructed in $O(n^3)$.

The generalized Spannogram algorithm for constructing $\mathcal{S}_d$ runs in $O(n^{d+1})$ and is formally presented in Alg. 2. A detailed description is provided in the supplemental material.

Algorithm 2 Spannogram algorithm for constructing $\mathcal{S}_d$

input $\mathbf{V} \in \mathbb{R}^{n\times d}$, $k \in [n]$.
$\mathcal{S}_d \leftarrow \{\}$ {Set of candidate supports in $\mathcal{R}(\mathbf{V})$}
$\widehat{\mathbf{V}} \leftarrow \begin{bmatrix}\mathbf{V}^T & \mathbf{0}_d\end{bmatrix}^T$ {Append zero row; $\widehat{\mathbf{V}} \in \mathbb{R}^{n+1\times d}$}
for all $\binom{n+1}{d}$ sets $\{i_1, \ldots, i_d\} \subseteq [n+1]$ **do**
  $\mathbf{c} \leftarrow \text{Nullspace}\left(\begin{bmatrix}\mathbf{e}_{i_1}^T - \mathbf{e}_{i_2}^T \\ \vdots \\ \mathbf{e}_{i_1}^T - \mathbf{e}_{i_d}^T\end{bmatrix}\widehat{\mathbf{V}} \in \mathbb{R}^{d-1\times d}\right)$
  for $\alpha \in \{+1, -1\}$ **do**
    $\mathcal{I} \leftarrow \mathcal{I}_k^+(\alpha\mathbf{V}\mathbf{c})$ {Entries $\geq$ than the k^{th} one}
    if $|\mathcal{I}| \leq k$ **then**
      $\mathcal{S}_d \leftarrow \mathcal{S}_d \cup \{\mathcal{I}\}$ {No ambiguity}
    else
      $\mathcal{A} \leftarrow \{i_1, \ldots, i_d\}\backslash\{n+1\}$ {Ambiguous set}
      $\mathcal{T} \leftarrow \mathcal{I}\backslash\{\{i_1, \ldots, i_d\} \cup \{n+1\}\}$
      $r \leftarrow k - |\mathcal{T}|$
      $\widehat{d} \leftarrow |\mathcal{A}|$
      for all $\binom{\widehat{d}}{r}$ r-subsets $\mathcal{A}^{(r)} \subseteq \mathcal{A}\}$ **do**
        $\widehat{\mathcal{I}} \leftarrow \mathcal{T} \cup \mathcal{A}^{(r)}$
        $\mathcal{S}_d \leftarrow \mathcal{S}_d \cup \{\widehat{\mathcal{I}}\}$ {$\leq 2^d$ new candidates}
      end for
    end if
  end for
end for
output $\mathcal{S}_d$

5. Quadratic maximization over unit vectors in the intersection of halfspaces

Each support set $\mathcal{I}$ in $\mathcal{S}_d$ yields a candidate nonnegative, k-sparse PC, which can be obtained by solving (16), a quadratic maximization over the intersection of halfspaces and the unit sphere:

$$\mathbf{c}_\star = \arg\max_{\substack{\mathbf{c}\in\mathbb{S}^d \\ \mathbf{R}\mathbf{c}\geq 0}} \mathbf{c}^T\mathbf{Q}\mathbf{c}, \qquad (P_d)$$

where $\mathbf{Q} = \mathbf{V}_{\mathcal{I}}^T\mathbf{V}_{\mathcal{I}}$ is a $d \times d$ matrix and $\mathbf{R}$ is a $k \times d$ matrix. Problem (P_d) is NP-hard: for $\mathbf{Q}$ PSD and $\mathbf{R} = \mathbf{I}_{d\times d}$, it reduces to the original problem in (2). Here, however, we are interested in the case where the dimension d is a constant. We outline an $O(k^d)$ algorithm, *i.e.*, polynomial in the number of linear constraints, for solving (P_d). A detailed proof is available in the supplemental material.

The objective of (P_d) is maximized by $\mathbf{u}_1 \in \mathbb{R}^d$, the leading eigenvector of $\mathbf{Q}$. If $\mathbf{u}_1$ or $-\mathbf{u}_1$ is feasible, *i.e.*, if it satisfies all linear constraints, then $\mathbf{c}_\star = \pm\mathbf{u}_1$. It can be shown that if none of $\pm\mathbf{u}_1$ is feasible, at least one of the k linear constraints is active at the optimal solution $\mathbf{c}_\star$, that is, there exists $1 \leq i \leq k$ such that $\mathbf{R}_{i,:}\mathbf{c}_\star = 0$.

Fig. 2 depicts an example for $d = 2$. The leading eigenvector of $\mathbf{Q}$ lies outside the feasible region, an arc of the unit-circle in the intersection of k halfspaces. The optimal solution coincides with one of the two endpoints of the feasible region, where a linear inequality is active, motivating a simple algorithm for solving (P_2): (*i*) for each linear inequality determine a unit length point where the inequality becomes active, and (*ii*) output the point that is feasible and maximizes the objective.

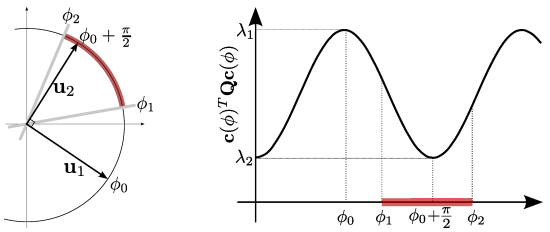


Figure 2. An instance of (P_2): the leading eigenvector $\mathbf{u}_1$ of $\mathbf{Q} \in \mathbb{S}^2$ lies outside the feasible region (highlighted arc). The maximum of the constrained quadratic optimization problem is attained at ϕ_1, an endpoint of the feasible interval.

Back to the general (P_d) problem, if a linear inequality $\mathbf{R}_{i,:}\mathbf{c} \geq 0$ for some $i \in [k]$ is enforced with equality, the modified problem can be written as a quadratic maximization in the form of (P_d), with dimension reduced to $d-1$ and $k-1$ linear constraints. This observation suggests a recursive algorithm for solving (P_d): If $\pm\mathbf{u}_1$ is feasible, it is also the optimal solution. Otherwise, for $i = 1, \ldots, k$, set the i^{th} inequality constraint active, solve recursively, and collect candidate solutions. Finally, output the candidate that maximizes the objective. The $O(k^d)$ recursive algorithm is formally presented in the supplemental material.

Algorithm 3 Approximate Spannogram NSPCA (ϵ-net)

input $\mathbf{A}$ ($n \times n$ PSD matrix), $k, d \in [n]$, $\epsilon \in (0, 1]$
$[\mathbf{U}, \mathbf{\Lambda}] = \texttt{svd}(\mathbf{A}, d)$
$\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$ $\{\mathbf{A}_d = \mathbf{V}\mathbf{V}^T\}$
$\mathcal{X}_d = \emptyset$
for $i = 1 : O(\epsilon^{-d} \cdot \log n)$ **do**
  $\mathbf{c} = \texttt{randn}(d, 1)$
  $\mathbf{a} = \mathbf{V}\mathbf{c}/\|\mathbf{c}\|_2$
  $\mathbf{x} = \texttt{rank1solver}(\mathbf{a})$ {Section 3.1}
  $\mathcal{X}_d = \mathcal{X}_d \cup \{\mathbf{x}\}$
end for
output $\widehat{\mathbf{x}}_d = \arg\max_{\mathbf{x} \in \mathcal{X}_d} \|\mathbf{V}^T\mathbf{x}\|_2^2$

6. Near-Linear Time Nonnegative SPCA

Alg. 1 approximates the nonnegative, k-sparse PC of a PSD matrix $\mathbf{A}$ by solving the nonnegative sparse PCA problem exactly on $\mathbf{A}_d$, the best rank-d approximation of $\mathbf{A}$. Albeit polynomial in n, the running time of Alg. 1 can be impractical even for moderate values of n.

Instead of pursuing the exact solution to the low-rank nonnegative sparse PCA problem $\max_{\mathbf{x} \in \mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_d\mathbf{x}$, we can compute an approximate solution in near-linear time, with performance arbitrarily close to optimal. The suggested procedure is outlined in Algorithm 3, and a detailed discussion is provided in the supplemental material. Alg. 3 relies on randomly sampling points from the range of $\mathbf{A}_d$ and efficiently solving rank-1 instances of the nonnegative sparse PCA problem as described in Section 3.1.

Theorem 2. *For any $n \times n$ PSD matrix $\mathbf{A}$, sparsity parameter k, and accuracy parameters $d \in [n]$ and $\epsilon \in (0, 1]$, Alg. 3 outputs a nonnegative, k-sparse, unit norm vector $\widehat{\mathbf{x}}_d$ such that*

$$\widehat{\mathbf{x}}_d^T\mathbf{A}\widehat{\mathbf{x}}_d \;\geq\; (1-\epsilon) \cdot \rho_d \cdot \mathbf{x}_\star{}^T\mathbf{A}\mathbf{x}_\star,$$

with probability at least $1 - 1/n$, in time $O\left(\epsilon^{-d} \cdot n \log n\right)$ plus the time to compute the d leading eigenvectors of $\mathbf{A}$.

7. Experimental Evaluation

We empirically evaluate the performance of our algorithm on various datasets and compare it to the EM algorithm[1] for sparse and nonnegative PCA of (Sigg & Buhmann, 2008) which is known to outperform previous algorithms.

CBCL Face Dataset. The CBCL face image dataset (Sung, 1996), with 2429 gray scale images of size 19×19 pixels, has been used in the performance evaluation of both the NSPCA (Zass & Shashua, 2007) and EM (Sigg & Buhmann, 2008) algorithms.

Fig. 3 depicts samples from the dataset, as well as six orthogonal, nonnegative, k-sparse components ($k = 40$) successively computed by *(i)* Alg. 3 ($d = 3, \epsilon = 0.1$) and *(ii)* the EM algorithm. Features active in one component are removed from the dataset prior to computing subsequent PCs to ensure orthogonality. Fig. 3 reveals the ability of nonnegative sparse PCA to extract significant parts.

In Fig. 4, we plot the variance explained by the computed approximate nonnegative, k-sparse PC (normalized by the leading eigenvalue) versus the sparsity parameter k. Alg. 3 for $d = 3$ and $\epsilon = 0.1$, and the EM algorithm exhibit nearly identical performance. For this dataset, we also compute the leading component using the NSPCA algorithm of (Zass & Shashua, 2007). Note that NSPCA does not allow for a precise control of the sparsity of its output; an appropriate sparsity penalty β was determined via binary search for each target sparsity k. We plot the explained variance only for those values of k for which a k-sparse component was successfully extracted. Finally, note that both the EM and NSPCA algorithms are randomly initialized. All depicted values are the best results over multiple random restarts.

Our theory allows us to obtain provable approximation guarantees: based on Theorem 2 and the output of Alg. 3, we compute a data dependent upper bound on the maximum variance, which provably lies in the shaded area. For instance, for $k = 180$, the extracted component explains at least 58% of the variance explained by the true nonnegative, k-sparse PC. The quality of the bound depends on the accuracy parameters d and ϵ, and the eigenvalue decay of the empirical covariance matrix of the data. There exist

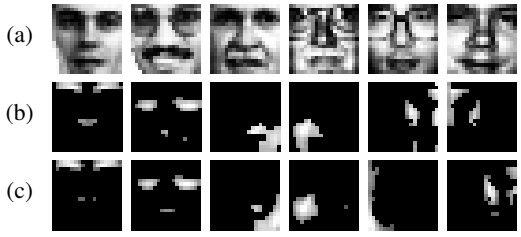


Figure 3. We plot (a) six samples from the dataset, and the six leading orthogonal, nonnegative, k-sparse PCs for $k = 40$ extracted by (b) Alg. 3 ($d = 3, \epsilon = 0.1$), and (c) the EM algorithm.

[1] Matlab implementation available by the author.

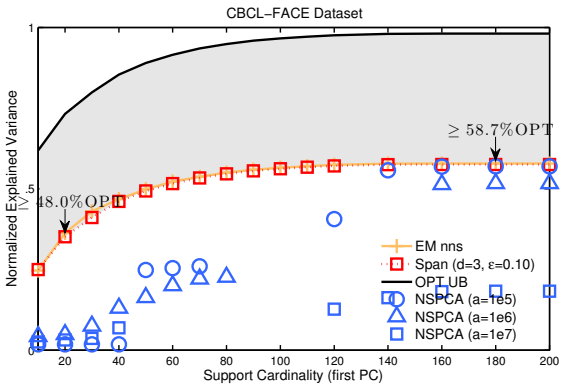


Figure 4. CBCL dataset (Sung, 1996). We plot the normalized variance explained by the approximate nonnegative, k-sparse PC versus the sparsity k. Our theory yields a provable data dependent approximation guarantee: the true unknown optimum provably lies in the shaded area.

datasets on which our algorithm provably achieves 70% or even 90% of the optimal.

Leukemia Dataset. The Leukemia dataset (Armstrong et al., 2001) contains 72 samples, each consisting of expression values for 12582 probe sets. The dataset was used in the evaluation of (Sigg & Buhmann, 2008). In Fig. 5, we plot the normalized variance explained by the computed nonnegative, k-sparse PC versus the sparsity parameter k. For low values of k, Alg. 3 outperforms the EM algorithm in terms of explained variance. For larger values, the two algorithms exhibit similar performance.

The approximation guarantees accompanying our algorithm allow us to upper bound the optimal performance. For k as small as 50, which roughly amounts to 0.4% of the features, the extracted component captures at least 44.6% of the variance corresponding to the true nonnegative k-sparse PC. The obtained upper bound is a significant improvement compared to the trivial bound given by λ_1.

Low Resolution Spectrometer Dataset. The Low Resolution Spectrometer (LRS) dataset, available in (Bache & Lichman, 2013), originates from the Infra-Red Astronomy Satellite Project. It contains 531 high quality spectra (samples) measured in 93 bands. Fig. 6 depicts the normalized variance explained by the computed nonnegative, k-sparse PC versus the sparsity parameter k. The empirical covariance matrix of this dataset exhibits sharper decay in the spectrum than the previous examples, yielding tighter approximation guarantees according to our theory. For instance, for $k = 20$, the extracted nonnegative component captures at least 86% of the maximum variance. For values closer to $k = 90$, where the computed PC is nonnegative but no longer sparse, this value climbs to nearly 93%.

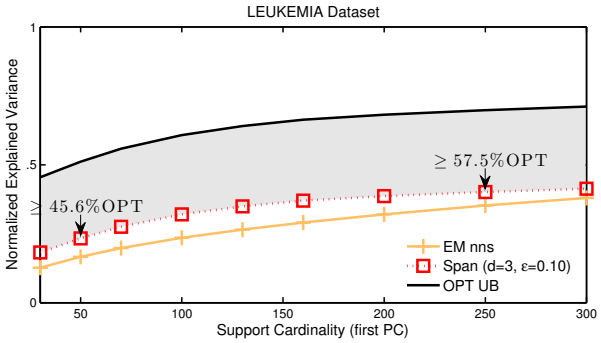


Figure 5. Leukemia dataset (Armstrong et al., 2001). We plot the normalized variance explained by the output of Alg. 3 ($d = 3, \epsilon = 0.1$) versus the sparsity k, and compare with the EM algorithm of (Sigg & Buhmann, 2008). By our approximation guarantees, the maximum variance provably lies in the shaded area.

8. Conclusions

We introduced a novel algorithm for nonnegative sparse PCA, expanding the spannogram theory to nonnegative quadratic optimization. We observe that the performance of our algorithm often matches and sometimes outperforms the previous state of the art (Sigg & Buhmann, 2008). Even though the theoretical running time of Alg. 3 scales better than EM, in practice we observed similar speed, both in the order of a few seconds. Our approach has the benefit of provable approximation, giving both theoretical a-priori guarantees and data dependent bounds that can be used to estimate the variance explained by nonnegative sparse PCs, as shown in our experiments.

9. Acknowledgements

The authors would like to acknowledge support from NSF grants CCF-1344364, CCF-1344179, DARPA XDATA and research gifts by Google and Docomo.

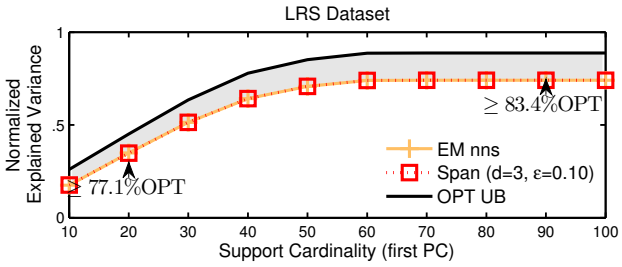


Figure 6. LRS dataset (Bache & Lichman, 2013). We plot the normalized explained variance versus the sparsity k. Alg. 3 ($d = 3, \epsilon = 0.1$) and the EM algorithm exhibit similar performance. The optimum value of the objective in (2) provably lies in the shaded area, which in this case is particularly tight.

References

Allen, Genevera I. and Maletić-Savatić, Mirjana. Sparse non-negative generalized pca with applications to metabolomics. *Bioinformatics*, 2011.

Armstrong, Scott A, Staunton, Jane E, Silverman, Lewis B, Pieters, Rob, den Boer, Monique L, Minden, Mark D, Sallan, Stephen E, Lander, Eric S, Golub, Todd R, and Korsmeyer, Stanley J. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, 30(1):41–47, 2001.

Asteris, M., Papailiopoulos, D.S., and Karystinos, G.N. Sparse principal component of a rank-deficient matrix. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pp. 673–677, 2011.

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Badea, Liviu and Tilivea, Doina. Sparse factorizations of gene expression guided by binding data. In *Pacific Symposium on Biocomputing*, 2005.

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

d'Aspremont, Alexandre, Bach, Francis R., and Ghaoui, Laurent El. Full regularization path for sparse principal component analysis. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pp. 177–184, New York, NY, USA, 2007b. ACM.

d'Aspremont, Alexandre, Bach, Francis, and Ghaoui, Laurent El. Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, 9:1269–1294, Jun 2008.

Hoyer, Patrik O. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5: 1457–1469, 2004.

Jolliffe, I.T. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22(1):29–35, 1995.

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

Karystinos, G.N. and Liavas, A.P. Efficient computation of the binary vector that maximizes a rank-deficient quadratic form. *Information Theory, IEEE Transactions on*, 56(7):3581–3593, 2010.

Kim, Hyunsoo and Park, Haesun. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12): 1495–1502, 2007.

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788–791, 1999.

Mackey, Lester. Deflation methods for sparse pca. In *Advances in Neural Information Processing Systems 21*, NIPS '08, pp. 1–8, Vancouver, Canada, Dec 2008.

Moghaddam, B., Weiss, Y., and Avidan, S. Spectral bounds for sparse pca: Exact and greedy algorithms. *Advances in neural information processing systems*, 18:915, 2006a.

Moghaddam, Baback, Weiss, Yair, and Avidan, Shai. Generalized spectral bounds for sparse lda. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pp. 641–648, 2006b.

Murty, Katta G. and Kabadi, Santosh N. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

Papailiopoulos, D. S., Dimakis, A. G., and Korokythakis, S. Sparse pca through low-rank approximations. In *Proceedings of the 30th International Conference on Machine Learning*, ICML '13, pp. 767–774. ACM, 2013.

Parrilo, Pablo A. *Structured Semidenite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, California Institute of Technology Pasadena, California, 2000.

Sigg, Christian D. and Buhmann, Joachim M. Expectation-maximization for sparse and non-negative pca. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 960–967, New York, NY, USA, 2008. ACM.

Sung, Kah-Kay. *Learning and example selection for object and pattern recognition*. PhD thesis, PhD thesis, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA, 1996.

Zass, Ron and Shashua, Amnon. Nonnegative sparse pca. In *Advances in Neural Information Processing Systems 19*, pp. 1561–1568, Cambridge, MA, 2007. MIT Press.

Zhang, Y., d'Aspremont, A., and Ghaoui, L.E. Sparse pca: Convex relaxations, algorithms and applications. *Handbook on Semidefinite, Conic and Polynomial Optimization*, pp. 915–940, 2012.

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

A. Approximation Guarantees

In this section, we develop a series of Lemmata that establish the approximation guarantees of Theorem 1. First, recall that

$$\mathsf{OPT} = \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}\mathbf{x}$$

corresponds to the optimal value of the quadratic objective function with argument $\mathbf{A}$, and let $\mathbf{x}_\star$ be the optimal solution, *i.e.*, the nonnegative, k-sparse, unit norm vector achieving value OPT. Similarly, OPT_d denotes the optimal value of the quadratic with argument $\mathbf{A}_d$, the best rank-d approximation of $\mathbf{A}$, that is

$$\mathsf{OPT}_d = \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_d\mathbf{x},$$

and $\mathbf{x}_d$ is the corresponding optimal solution. Alg. 1 with input $\mathbf{A}$ and accuracy parameter d computes and outputs $\mathbf{x}_d$ as a surrogate for the desired vector $\mathbf{x}_\star$. We show that

$$\mathbf{x}_d{}^T\mathbf{A}\mathbf{x}_d \geq \rho_d \cdot \mathbf{x}_\star{}^T\mathbf{A}\mathbf{x}_\star,$$

where

$$\rho_d \geq \max\left\{\frac{k}{2n},\ \frac{1}{1+2\frac{n}{k}\lambda_{d+1}/\lambda_1}\right\}.$$

Lemma A.1. *Let $\mathbf{x}_d$ denote the nonnegative k-sparse principal component of $\mathbf{A}_d$, i.e., $\mathbf{x}_d = \arg\max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_d\mathbf{x}$, achieving value $\mathsf{OPT}_d = \mathbf{x}_d{}^T\mathbf{A}_d\mathbf{x}_d$. Then,*

$$\mathbf{x}_d{}^T\mathbf{A}\mathbf{x}_d \ \geq\ \mathsf{OPT}_d.$$

Proof. The lemma is a consequence of the fact that $\mathbf{A}$ is a positive semidefinite matrix:

$$\begin{aligned}\mathbf{x}_d{}^T\mathbf{A}\mathbf{x}_d &= \mathbf{x}_d{}^T\left(\sum_{i=1}^n \lambda_i\mathbf{q}_i\mathbf{q}_i^T\right)\mathbf{x}_d\\ &= \mathbf{x}_d{}^T\mathbf{A}_d\mathbf{x}_d + \sum_{i=d+1}^n \left|\sqrt{\lambda_i}\mathbf{q}_i^T\mathbf{x}_d\right|^2\\ &\geq \mathsf{OPT}_d, \quad \forall d\in[n],\end{aligned}$$

which is the desired result. □

Lemma A.2. *The optimal value OPT_d of the rank-d nonnegative k-sparse PCA problem satisfies*

$$\mathsf{OPT} - \lambda_{d+1} \ \leq\ \mathsf{OPT}_d \ \leq\ \mathsf{OPT}.$$

Proof. The upper bound is due to the fact that $\mathbf{A}$ is a positive semidefinite matrix:

$$\begin{aligned}\mathsf{OPT} &= \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}\mathbf{x}\\ &\geq \mathbf{x}_d{}^T\mathbf{A}\mathbf{x}_d\\ &= \mathbf{x}_d{}^T\mathbf{A}_d\mathbf{x}_d + \mathbf{x}_d{}^T\left(\mathbf{A}-\mathbf{A}_d\right)\mathbf{x}_d\\ &= \mathsf{OPT}_d + \sum_{i=d+1}^n \lambda_i|\mathbf{q}_i{}^T\mathbf{x}_d|^2\\ &\geq \mathsf{OPT}_d.\end{aligned}$$

For the lower bound,

$$\begin{aligned}\mathsf{OPT} &= \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}\mathbf{x}\\ &= \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\left(\sum_{i=1}^n \lambda_i\mathbf{q}_i\mathbf{q}_i^T\right)\mathbf{x}\\ &\leq \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_d\mathbf{x} + \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\sum_{i=d+1}^n \lambda_i\mathbf{q}_i\mathbf{q}_i^T\mathbf{x}\\ &\leq \mathsf{OPT}_d + \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\sum_{i=d+1}^n \lambda_i\mathbf{q}_i\mathbf{q}_i^T\mathbf{x}\\ &\leq \mathsf{OPT}_d + \max_{\|\mathbf{x}\|=1} \mathbf{x}^T\sum_{i=d+1}^n \lambda_i\mathbf{q}_i\mathbf{q}_i^T\mathbf{x}\\ &= \mathsf{OPT}_d + \lambda_{d+1},\end{aligned}$$

which completes the proof. □

Lemma A.3.

$$\frac{\mathsf{OPT}_d}{\mathsf{OPT}} \ \geq\ \max\left\{\frac{\mathsf{OPT}_d}{\lambda_1},\ \frac{1}{1+\frac{\lambda_{d+1}}{\mathsf{OPT}_d}}\right\},\ \forall d\in[n].$$

Proof. It suffices to show that $\mathsf{OPT}_d/\mathsf{OPT}$ is lower bounded by both quantities on the right-hand side. The first lower bound follows trivially from the fact that

$$\mathsf{OPT} = \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}\mathbf{x} \ \leq\ \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_1.$$

For the second lower bound, note that by Lemma A.2, $\mathsf{OPT} \leq \mathsf{OPT}_d + \lambda_{d+1}$, which in turn implies

$$\frac{\mathsf{OPT}_d}{\mathsf{OPT}} \ \geq\ \frac{1}{1+\lambda_{d+1}/\mathsf{OPT}_d}.$$

□

Lemma A.4. *The optimal value OPT_1 of the nonnegative, k-sparse PCA problem $\max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_1\mathbf{x}$ on the rank-1 matrix $\mathbf{A}_1$ satisfies*

$$\mathsf{OPT}_1 \ \geq\ \frac{1}{2}\frac{k}{n}\lambda_1.$$

Proof. Let $(\mathbf{v})_k^+$ denote the vector obtained by setting to zero all but the (at most) k largest nonnegative entries of $\mathbf{v}$. By definition,

$$\begin{aligned}\mathsf{OPT}_1 &= \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_1\mathbf{x}\\ &= \lambda_1 \cdot \max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{q}_1\mathbf{q}_1^T\mathbf{x}\\ &= \lambda_1 \cdot \max_{\mathbf{x}\in\mathbb{S}_k^n} \left|\mathbf{q}_1^T\mathbf{x}\right|^2\\ &= \lambda_1 \cdot \max\left\{\left|\frac{\mathbf{q}_1^T(\mathbf{q}_1)_k^+}{\|(\mathbf{q}_1)_k^+\|}\right|^2, \left|\frac{\mathbf{q}_1^T(-\mathbf{q}_1)_k^+}{\|(-\mathbf{q}_1)_k^+\|}\right|^2\right\}\\ &= \lambda_1 \cdot \max\left\{\|(\mathbf{q}_1)_k^+\|^2,\ \|(-\mathbf{q}_1)_k^+\|^2\right\}.\end{aligned}$$

It holds that

$$\|(\mathbf{q}_1)_k^+\|^2 \ \geq\ \frac{k}{n}\|(\mathbf{q}_1)_n^+\|^2. \tag{17}$$

To verify that, let $\mathcal{I}_k$ be the support of $(\mathbf{q}_1)_k^+$ and $\mathcal{I}_n$ the support of $(\mathbf{q}_1)_n^+$. Clearly, $\mathcal{I}_k \subseteq \mathcal{I}_n$. Let u be the value of the smallest non-zero entry in $(\mathbf{q}_1)_k^+$. This implies that

$$\|(\mathbf{q}_1)_k^+\|^2 = \sum_{i\in\mathcal{I}_k}([\mathbf{q}_1]_i)^2 \ \geq\ k\cdot u^2.$$

Further,

$$\begin{aligned}\|(\mathbf{q}_1)_n^+\|^2 &= \sum_{i\in\mathcal{I}_k}([\mathbf{q}_1]_i)^2 + \sum_{i\in\mathcal{I}_n\setminus\mathcal{I}_n}([\mathbf{q}_1]_i)^2\\ &= \|(\mathbf{q}_1)_k^+\|^2 + \sum_{i\in\mathcal{I}_n\setminus\mathcal{I}_k}([\mathbf{q}_1]_i)^2\\ &\leq \|(\mathbf{q}_1)_k^+\|^2 + (n-k)\cdot u^2,\end{aligned}$$

From the two inequalities, it follows that

$$\frac{\|(\mathbf{q}_1)_n^+\|^2}{\|(\mathbf{q}_1)_k^+\|^2} \leq 1 + \frac{(n-k)\cdot u^2}{k\cdot u^2} \leq \frac{n}{k},$$

which in turn implies (17). By the same argument,

$$\|(-\mathbf{q}_1)_k^+\|^2 \geq \frac{k}{n}\|(-\mathbf{q}_1)_n^+\|^2. \tag{18}$$

Finally, noting that

$$1 = \|\mathbf{q}_1\|^2 = \|(\mathbf{q}_1)_n^+\|^2 + \|(-\mathbf{q}_1)_n^+\|^2,$$

and combining with (17) and (18), we obtain

$$\begin{aligned}\mathsf{OPT}_1 &\geq \lambda_1\cdot\max\left\{\frac{k}{n}\|(\mathbf{q}_1)_n^+\|^2,\ \frac{k}{n}\|(-\mathbf{q}_1)_n^+\|^2\right\}\\ &= \frac{k}{n}\lambda_1\cdot\max\left\{\|(\mathbf{q}_1)_n^+\|^2,\ 1-\|(\mathbf{q}_1)_n^+\|^2\right\}\\ &\geq \frac{1}{2}\frac{k}{n}\lambda_1,\end{aligned}$$

which completes the proof. □

Lemma A.5.

$$\frac{\mathsf{OPT}_d}{\mathsf{OPT}} \ \geq\ \max\left\{\frac{k}{2n},\ \frac{1}{1+2\frac{n}{k}\lambda_{d+1}/\lambda_1}\right\}.$$

Proof. $\mathbf{A}_1$ is the best rank-1 approximation of $\mathbf{A}_d$. By Lemmata A.2 and A.4, we have $\mathsf{OPT}_d \geq \mathsf{OPT}_1 \geq \frac{1}{2}\frac{k}{n}\lambda_1$, for all $d \geq 1$. The desired result follows from Lemma A.3 and the previous lower bound on OPT_d. □

Proof of Theorem 1. By Lemma A.1, ${\mathbf{x}_d}^T\mathbf{A}\mathbf{x}_d \geq \mathsf{OPT}_d$. Dividing both sides by $\mathsf{OPT} = {\mathbf{x}_\star}^T\mathbf{A}\mathbf{x}_\star$, we obtain

$${\mathbf{x}_d}^T\mathbf{A}\mathbf{x}_d \ \geq\ \rho_d\cdot{\mathbf{x}_\star}^T\mathbf{A}\mathbf{x}_\star,$$

where $\rho_d = \mathsf{OPT}_d/\mathsf{OPT}$. The lower bound on ρ_d given in Theorem 1 follows from Lemma A.5. The computational complexity of Alg. 1 follows from the detailed description of the algorithm and is analyzed separately. ■

A.1. Approximation Guarantees - Special Cases

Corollary 1. *If the eigenvalues of* $\mathbf{A}$ *follow a decay law* $\lambda_i \leq \lambda_1\cdot f(i)$ *for any vanishing function* $f(i)$*, i.e., for* $f(i)\to 0$ *as* $i\to\infty$*, then for* $k = c\cdot n$*, where* c *is constant* $0 < c \leq 1$ *(linear sparsity regime), Alg. 1 yields a polynomial time approximation scheme (PTAS). That is, for any constant* ϵ*, we can choose a constant accuracy parameter* d *and obtain a solution* $\mathbf{x}_d$ *such that*

$${\mathbf{x}_d}^T\mathbf{A}\mathbf{x}_d \ \geq\ (1-\epsilon)\cdot\mathsf{OPT},$$

in time polynomial in n *and* k*, but not in* $1/\epsilon$*.*

Proof. By assumption, $\lambda_i \leq \lambda_1\cdot f(i)$ for some function $f(i)$ such that $f(i)\to 0$ as $i\to\infty$. For any constants c and ϵ, there must hence exists a finite i such that

$$f(i) \leq \frac{c}{2}\cdot\frac{\epsilon}{1-\epsilon}.$$

Set d equal to the smallest i for which the above holds: d will be some function $g(\epsilon)$ that depends on $f(\cdot)$. By Theorem 1, and under the assumption of the corollary we have

$$\begin{aligned}\rho_d \ &\geq\ \frac{1}{1+2\frac{n}{k}\lambda_{d+1}/\lambda_1} \ \geq\ \frac{1}{1+2f(d)/c}\\ &\geq\ \frac{1}{1+\epsilon/(1-\epsilon)} \ \geq\ (1-\epsilon),\end{aligned}$$

which implies that for any ϵ, the output $\mathbf{x}_d$ will be within factor $1-\epsilon$ from the optimal. Alg. 1 runs in time $O(n^{2d}) = O(n^{g(\epsilon)})$, which completes the proof. □

B. The Spannogram Algorithm

In this section, we provide a detailed description of the Spannogram algorithm for the construction of the collection $\mathcal{S}_d$ of candidate support sets in the case of arbitrary d.

For completeness, we first give a proof for Proposition 3.1, which states that the support of $\mathbf{x}_1$, the optimal solution of the rank-1 nonnegative sparse PCA problem

$$\max_{\mathbf{x}\in\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_1\mathbf{x} = \max_{\mathbf{x}\in\mathbb{S}_k^n} \left(\mathbf{v}^T\mathbf{x}\right)^2,$$

where $\mathbf{A}_1 = \mathbf{v}\mathbf{v}^T$, coincides with one of the two sets in the collection $\mathcal{S}_1 = \left\{\mathcal{I}_k^+(\mathbf{v})\,,\ \mathcal{I}_k^+(-\mathbf{v})\right\}$.

B.1. Proof of Proposition 3.1

Let $\mathbf{x}_\star = \arg\max_{\mathbf{x}\in\mathbb{S}_k^n} \left(\mathbf{v}^T\mathbf{x}\right)^2$. First, assume that $\mathbf{v}^T\mathbf{x}_\star \geq 0$. We will show that $\text{supp}\,(\mathbf{x}_1^\star) \subseteq \mathcal{I}_k^+(\mathbf{v})$.

Assume, for the sake of contradiction, that the support of $\mathbf{x}_\star$ does *not* coincide with $\mathcal{I}_k^+(\mathbf{v})$. This implies that there exists an index $j \notin \mathcal{I}_k^+(\mathbf{v})$ such that $j \in \text{supp}\,(\mathbf{x}_\star)$, *i.e.*, $[\mathbf{x}_\star]_j > 0$. By the definition of $\mathcal{I}_k^+(\mathbf{v})$, $j \notin \mathcal{I}_k^+(\mathbf{v})$ implies that either *(i)* $v_j < 0$, or *(ii)* there exist at least k nonnegative entries in $\mathbf{v}$ larger than v_j.

In the first case, consider a vector $\widehat{\mathbf{x}}$ that is equal to $\mathbf{x}_\star$ in all entries except the j^{th} entry which is set to zero in $\widehat{\mathbf{x}}$. Then, $\mathbf{y} = \widehat{\mathbf{x}}/\|\widehat{\mathbf{x}}\|_2$, is k-sparse (at most $k-1$ nonzero entries), nonnegative and unit length. It should not be hard to see that since $v_j < 0$, $\mathbf{v}^T\mathbf{y} \geq \mathbf{v}^T\mathbf{x}_\star$, contradicting the optimality of $\mathbf{x}_\star$.

In the second case, let l be the index of one of the k largest nonnegative entries in $\mathbf{v}$ such that $[\mathbf{x}_\star]_l = 0$. Such an entry exists, because otherwise $\mathbf{x}_\star$ would have more than k nonzero entries. Construct a nonnegative, k-sparse, unit length vector $\mathbf{y}$ by swapping the values in the j^{th} and l-th entries of $\mathbf{x}_\star$. Then, $\mathbf{v}^T\mathbf{y} \geq \mathbf{v}^T\mathbf{x}_\star$, contradicting the optimality of $\mathbf{x}_\star$.

We conclude that

$$\mathbf{v}^T\mathbf{x}_\star \geq 0 \quad \Rightarrow \quad \text{supp}(\mathbf{x}_\star) \subseteq \mathcal{I}_k^+(\mathbf{v})\,.$$

Similarly, if $\mathbf{v}^T\mathbf{x}_\star < 0$, then $-\mathbf{v}^T\mathbf{x}_\star > 0$, and

$$\mathbf{v}^T\mathbf{x}_\star < 0 \quad \Rightarrow \quad \text{supp}(\mathbf{x}_\star) \subseteq \mathcal{I}_k^+(-\mathbf{v})\,.$$

Since either $\mathbf{v}^T\mathbf{x}_\star \geq 0$ or $\mathbf{v}^T\mathbf{x}_\star < 0$ holds, we conclude that $\text{supp}(\mathbf{x}_\star) \in \mathcal{S}_1 = \left\{\mathcal{I}_k^+(\mathbf{v})\,,\ \mathcal{I}_k^+(-\mathbf{v})\right\}$. ■

B.2. The general rank-d case

We generalize the developments of Section 4.1 to case of arbitrary constant d. More specifically, we will show that for any $\mathbf{V} \in \mathbb{R}^{n\times d}$, the collection

$$\mathcal{S}_d = \bigcup_{\mathbf{c}\in\mathbb{S}^d} \left\{\mathcal{I}_k^+(\mathbf{V}\mathbf{c})\right\}$$

contains at most $O(n^d)$ candidate support sets and can be constructed in $O(n^{d+1})$.

Hyperspherical variables. Let $\mathcal{R}(\mathbf{V})$ denote the range of $\mathbf{V} \in \mathbb{R}^{n\times d}$. Up to scaling, all vectors $\mathbf{v}$ in $\mathcal{R}(\mathbf{V})$, can be written as $\mathbf{v} = \mathbf{V}\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^d : \|\mathbf{c}\| = 1$. We introduce $d-1$ variables $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_{d-1}] \in \Phi^{d-1} = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right]^{d-1}$, and set $\mathbf{c}$ to be the following function of $\boldsymbol{\phi}$:

$$\mathbf{c}(\boldsymbol{\phi}) = \begin{bmatrix} \sin(\phi_1) \\ \cos(\phi_1)\sin(\phi_2) \\ \vdots \\ \cos(\phi_1)\cos(\phi_2)\cdots\sin(\phi_{d-1}) \\ \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{d-1}) \end{bmatrix} \in \mathbb{R}^d. \tag{19}$$

In other words, $\phi_1, \ldots, \phi_{d-1}$ are the spherical coordinates of $\mathbf{c}(\boldsymbol{\phi})$. All unit vectors in $\mathbb{R}^d$ can be mapped to a spherical coordinate vector $\boldsymbol{\phi} \in (-\pi, \pi]\times\Phi^{d-2}$. Restricting variable ϕ_1 to Φ limits $\mathbf{c}(\boldsymbol{\phi})$ to half the d-dimensional unit sphere: for any unit norm vector $\mathbf{c}$, there exists $\boldsymbol{\phi} \in \Phi^{d-1}$ such that $\mathbf{c} = \mathbf{c}(\boldsymbol{\phi})$ or $\mathbf{c} = -\mathbf{c}(\boldsymbol{\phi})$.

Under (19), the vectors in $\mathcal{R}(\mathbf{V})$ can be described as a function of $\boldsymbol{\phi}$: $\mathcal{R}(\mathbf{V}) = \{\pm\mathbf{v}(\boldsymbol{\phi}) = \pm\mathbf{V}\mathbf{c}(\boldsymbol{\phi}), \boldsymbol{\phi} \in \Phi^{d-1}\}$. In turn, the set of indices of the k largest nonnegative entries of $\mathbf{v}(\boldsymbol{\phi})$ is itself a function of $\boldsymbol{\phi}$, and

$$\mathcal{S}_d = \bigcup_{\boldsymbol{\phi}\in\Phi^{d-1}} \left\{\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))\,,\ \mathcal{I}_k^+(-\mathbf{v}(\boldsymbol{\phi}))\right\}.$$

Spannogram. The i^{th} entry of $\mathbf{v}(\boldsymbol{\phi})$ is

$$[\mathbf{v}(\boldsymbol{\phi})]_i = V_{i,1}\sin(\phi_1) + \cdots + V_{i,d}\prod_{l=1}^{d}\cos(\phi_{i_l}),$$

a continuous function of $\boldsymbol{\phi} \in \Phi^{d-1}$; a $(d-1)$-dimensional hypersurface in the d-dimensional space $\Phi^{d-1}\times\mathbb{R}$, for all $i \in [n]$. The collection of hypersurfaces constitutes the rank-d spannogram. As an example, Fig. 7 depicts the spannogram of an arbitrary 4×3 ($d = 3$) matrix $\mathbf{V}$.

At any particular point $\boldsymbol{\phi} \in \Phi^{d-1}$, assuming that no two hypersurfaces intersect at $\boldsymbol{\phi}$, the set $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ can be readily determined: sort the entries of $\mathbf{v}(\boldsymbol{\phi})$ and pick the indices of the at most k largest nonnegative entries. Note, however, that constructing $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ does not require a complete sorting the entries of $\mathbf{v}(\boldsymbol{\phi})$: detecting the k^{th} order entry and the (at most) $k-1$ nonnegative entries larger than that can be done in $O(n)$.

The key observation of our algorithm is that, due to their continuity, the hypersurfaces will retain their sorting

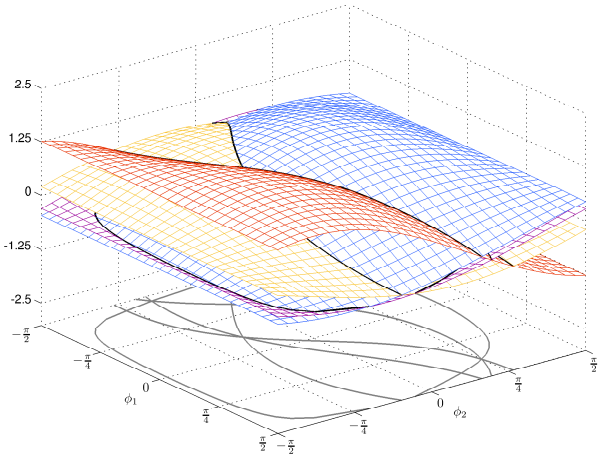


Figure 7. Spannogram of an arbitrary rank-3 matrix $\mathbf{V} \in \mathbb{R}^{4\times 3}$. Every surface is generated by one row of $\mathbf{V}$. At every point $\boldsymbol{\phi} = [\phi_1, \phi_2]$, the surface values correspond to the entries of a vector in the range of $\mathbf{V}$.

around $\boldsymbol{\phi}$ and hence, $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ tends to remain invariant. Moving away from $\boldsymbol{\phi}$, $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ can only change if when either the sign of a hypersurface or its order relative to other hypersurfaces changes. In other words, $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ can only change at points $\boldsymbol{\phi} \in \Phi^{d-1}$ where *(i)* two hypersurfaces intersect, or *(ii)* a hypersurface crosses the zero-hypersurface. Henceforth, we will assume that $\mathbf{V}$ has $n+1$ rows, where the last row is the zero vector, $\mathbf{0}_d$, generating the zero-hypersurface. As a result, the points of interest lie in the intersection of subsets of the $n+1$ hypersurfaces in the spannogram of $\mathbf{V}$.

We have argued that in order to construct $\mathcal{S}_d$, it suffices to consider points corresponding to the intersection of pairs of hypersurfaces. For $d > 2$, pairwise hypersurface intersections no longer correspond to single points. In the sequel, however, we will show that the points of interest can be further reduced to a finite set of points.

Let us examine when the set $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ changes from the perspective of the i^{th} hypersurface. That is, we ask what are the points in Φ^{d-1} where the i^{th} index might join or leave the candidate support set $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$. We know it suffices to examine only those points in Φ^{d-1} at which the i^{th} hypersurface intersects with another of the $n+1$ hypersurfaces. Let us focus on the intersection with the j^{th} hypersurface, $j \in [n+1], j \neq i$. We define

$$\mathcal{H}(i,j) = \left\{\mathbf{v}(\boldsymbol{\phi}) : [\mathbf{v}(\boldsymbol{\phi})]_i = [\mathbf{v}(\boldsymbol{\phi})]_j, \boldsymbol{\phi} \in \Phi^{d-1}\right\},$$

as the set of points lying in the intersection of hypersurfaces i and j. These points form a $(d-2)$-dimensional hypersurface[2]. Further, let

$$\Phi(i,j) = \{\boldsymbol{\phi} : \mathbf{v}(\boldsymbol{\phi}) \in \mathcal{H}(i,j)\},$$

be the corresponding $\boldsymbol{\phi}$'s. By definition, at every $\boldsymbol{\phi} \in \Phi(i,j)$, hypersurfaces i and j have the same values, and in opposite directions over $\boldsymbol{\phi} \in \Phi(i,j)$ the relative order of the two hypersurfaces changes. However, not all of the points in $\Phi(i,j)$ are necessarily points of interest; it is not necessary that $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ changes at every $\boldsymbol{\phi} \in \Phi(i,j)$. We seek to restrict our attention to a smaller subset of points.

If at some $\boldsymbol{\phi} \in \Phi(i,j)$ the i^{th} hypersurface is included or excluded from $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$, we ask what are those points where index i might leave or join the candidate support set. Once again, due to the continuity of the hypersurfaces, the set $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ is locally invariant as we scan $\Phi(i,j)$. The points of interest are those points at which the i^{th} hypersurface intersects another hypersurface. Provided that $\Phi(i,j)$ corresponds to points where the i^{th} and j^{th} hypersurfaces coincide, any intersection of the i^{th} hypersurface with a third hypersurface, say the l-th one, will be a joint intersection of the three hypersurfaces $\{i,j,l\}$. The set of points where the hypersurfaces $\{i,j,l\}$ intersect is

$$\mathcal{H}(i,j,l) \subseteq \mathcal{H}(i,j),$$

for all $l \in [n+1]\backslash\{i,j\}$. Repeating this argument recursively, we conclude that it suffices to examine the intersections of subsets of d hypersurfaces, $\mathcal{H}(i_1,i_2,\ldots,i_d)$, for all possible sets $\{i_1,i_2,\ldots,i_d\} \subseteq [n+1]$. Such intersections correspond to single points[3], where d hypersurfaces have the same value. By our perturbation argument, we can assume that exactly (*i.e.*, not more than) d hypersurfaces intersects at that exact $\boldsymbol{\phi}$. If all d intersecting hypersurfaces or none of them are included or excluded from $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$, the candidate set does not change (at least from the perspective of the i^{th} hypersurface) around $\boldsymbol{\phi}$. That is, index i neither leaves nor joins the candidate support set at $\boldsymbol{\phi}$. On the contrary, if the d intersecting hypersurfaces $\{i_1,i_2,\ldots,i_d\}$ are nonnegative and in the k-th order at $\boldsymbol{\phi}$, then there are multiple candidate support sets associated with the area around $\boldsymbol{\phi}$. In each of these candidates, only a subset of $\{i_1,i_2,\ldots,i_d\}$ can be included in $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$, due to the constraint that $|\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))| \leq k$. However, hypersurfaces $\{i_1,i_2,\ldots,i_d\}$ are the only ones that might join or leave the candidate set at that particular point and there are at most 2^{d-1} partitions of $\{i_1,i_2,\ldots,i_d\}$ into two subsets. Hence, at most a constant number of candidates, readily determined, is associated with each such intersection point.

[2] In the rank-2 case, the intersection was a single point.

[3] We assume that every d rows of $\mathbf{V}$ are linearly independent. If that is not the case, we can ignore the

Building $\mathcal{S}_d$. We consider all points where d hypersurfaces intersect, *i.e.*, we find $\boldsymbol{\phi}$ such that

$$[\mathbf{v}(\boldsymbol{\phi})]_{i_1} = [\mathbf{v}(\boldsymbol{\phi})]_{i_2} = \ldots = [\mathbf{v}(\boldsymbol{\phi})]_{i_d},$$

for all possible sets $\{i_1, \ldots, i_d\} \subseteq [n+1]$. To that end, it suffices to find $\boldsymbol{\phi}$ where the pairwise equalities

$$[\mathbf{v}(\boldsymbol{\phi})]_{i_1} = [\mathbf{v}(\boldsymbol{\phi})]_{i_2},\ \ldots\ , [\mathbf{v}(\boldsymbol{\phi})]_{i_1} = [\mathbf{v}(\boldsymbol{\phi})]_{i_d}$$

are jointly satisfied, or, equivalently, to find $\mathbf{c}(\boldsymbol{\phi})$ such that

$$\begin{bmatrix} \mathbf{e}_{i_1}^T - \mathbf{e}_{i_2}^T \\ \vdots \\ \mathbf{e}_{i_1}^T - \mathbf{e}_{i_{d-1}}^T \end{bmatrix} \mathbf{V}\mathbf{c}(\boldsymbol{\phi}) = \mathbf{0}_{d-1}.$$

In other words, we seek the unique (up to scaling) vector in the nullspace of the $d-1 \times d$ matrix multiplying $\mathbf{c}(\boldsymbol{\phi})$.

At the intersection point, the hypersurfaces indexed by $\{i_1, \ldots, i_d\}$ are all equal. If all d intersecting hypersurfaces are all (or none) included in $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$, then any modification in their sorting does not affect the set, in the sense that none of these d hypersurfaces leaves or joins the set of k largest nonnegative hypersurfaces. On the other hand, if the d hypersurfaces are nonnegative and equal to the k^{th} order hypersurface, then only a subset of them can be included in $\mathcal{I}_k^+(\mathbf{v}(\boldsymbol{\phi}))$ at any point around the intersection point. Further these are the only hypersurfaces that can leave or join the set at that point. If $1 \leq r \leq d-1$ of them can be included, then there must be at most $\binom{d}{r}$ candidates associated with the cells around $\boldsymbol{\phi}$ (or $\binom{d-1}{r}$ if one of them is the artificial zero hypersurface). That is, there can be at most $\binom{d}{\lceil \frac{d}{2} \rceil} \leq 2^{d-1}$ candidate support sets around $\boldsymbol{\phi}$.

We repeat the process for $\mathcal{I}_k^+(-\mathbf{v}(\boldsymbol{\phi}))$. Therefore, a maximum of $2 \cdot 2^{d-1}$ candidates are introduced at each intersection point, and

$$|\mathcal{S}_d| \leq 2^d \binom{n+1}{d} = O(n^d).$$

The candidates at each intersection point are determined in linear time: determining the entries of $\mathbf{v}(\boldsymbol{\phi})$ that are greater than its k-th largest nonnegative entry can be done in linear time, and the algorithm produces at most 2^{d-1} candidates at each intersection point. We conclude that $\mathcal{S}_d$ can be constructed in $O(n^{d+1})$.

C. Quadratic maximization over unit length vectors in the intersection of halfspaces

We consider the constrained quadratic maximization

$$\mathbf{c}_\star = \arg \max_{\substack{\|\mathbf{c}\|_2 = 1, \\ \mathbf{R}\mathbf{c} \geq 0}} \mathbf{c}^T \mathbf{Q} \mathbf{c}, \qquad (P_d)$$

where $\mathbf{Q}$ is a $d \times d$ symmetric matrix, and $\mathbf{R}$ is a real $k \times d$ matrix. In general, (P_d) is NP-hard: for $\mathbf{Q}$ PSD and $\mathbf{R}$ equal to the identity matrix $\mathbf{I}_d$, (P_d) reduces to the original problem in (2). In this section, however, we consider the case where the dimension d of the problem is a constant and develop an $O(k^d)$ algorithm for the non-trivial task of solving (P_d).

The $d = 1$ case. The optimization variable $\mathbf{c}$ is a scalar in $\{+1, -1\}$, and $\mathbf{R}$ is a vector in $\mathbb{R}^k$. If either $\mathbf{R} \geq \mathbf{0}$ or $-\mathbf{R} \geq \mathbf{0}$, the optimal solution is $\mathbf{c}^\star = 1$ or -1, respectively. Otherwise, the problem is infeasible.

The $d = 2$ case. The $d = 2$ case is the simplest nontrivial case. Let $\lambda_1 \geq \lambda_2$, be the two eigenvalues of $\mathbf{Q} \in \mathbb{S}^2$. The corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2$ form an orthonormal basis of $\mathbb{R}^2$. Any unit length vector $\mathbf{c}$ can be expressed as $\mathbf{c}(\phi) = \mathbf{U}\left[\cos(\phi),\ \sin(\phi)\right]^T$, for some $\phi \in [0, 2\pi)$, where $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2]$.

The feasible region is an arc on the unit circle in the intersection of k half-spaces (see Fig. 2 for an example). It comprises vectors $\mathbf{c}(\phi)$ with ϕ restricted in some interval $[\phi_1,\ \phi_2]$. Note that ϕ_1 and ϕ_2 are points where at least one linear constraint becomes active. Unless $\mathbf{R}$ is the zero matrix, $0 \leq |\phi_1 - \phi_2| \leq \pi$. If $\pm\mathbf{u}_1$ lies in the feasible region, then $\mathbf{c}_\star = \pm\mathbf{u}_1$: the leading eigenvector is the global unconstrained maximum. The key observation is that if neither $\mathbf{u}_1$ or $-\mathbf{u}_1$ is feasible, the optimal solution coincides with either $\mathbf{c}(\phi_1)$ or $\mathbf{c}(\phi_2)$. To verify that, let

$$Q(\phi) = \mathbf{c}(\phi)^T \mathbf{Q} \mathbf{c}(\phi) = \cos^2(\phi)\lambda_1 + \sin^2(\phi)\lambda_2$$

denote the quadratic objective in (P_2) as a function of ϕ. $Q(\phi)$ is differentiable with four critical points at $\phi = 0, \pi/2, \pi$, and $3\pi/2$. By assumption, $\phi = 0$ and $\phi = \pi$, which correspond to $\mathbf{c}(\phi) \pm \mathbf{u}_1$, lie outside the feasible interval $[\phi_1,\ \phi_2]$. Since $0 \leq |\phi_1 - \phi_2| \leq \pi$, at most one of the local minima $\phi = \pi/2$ and $\phi = 3\pi/2$ may lie in $[\phi_1, \phi_2]$. We conclude that either *(i)* $Q(\phi)$ is monotonically increasing in $[\phi_1, \phi_2]$, *(ii)* monotonically decreasing in $[\phi_1, \phi_2]$, or *(iii)* has a unique local minimum in (ϕ_1, ϕ_2). In either case, $Q(\phi)$ attains its maximum at one ϕ_1 and ϕ_2.

The above motivate the following steps for solving (P_2):

1. If $\pm\mathbf{R}\mathbf{u}_1 \geq \mathbf{0}$, then $\mathbf{c}_\star = \pm\mathbf{u}_1$.
2. Otherwise, initialize an empty collection $\mathcal{C}$ of candidate solutions. For $i = 1, \ldots, k$:
   - Compute $\mathbf{c}_i = \pm\left[-R_{i,2},\ R_{i,1}\right]^T / \|\mathbf{R}_{i,:}\|$, the unit norm vectors in the direction at which the i^{th} inequality is active. If $\pm\mathbf{c}_i$ is feasible, include $\pm\mathbf{c}_i$ in $\mathcal{C}$.
3. Return $\mathbf{c}_\star = \arg\max_{\mathbf{c} \in \mathcal{C}} \mathbf{c}^T \mathbf{Q} \mathbf{c}$.

The previous steps are formally presented in Algorithm 4.

Lemma C.6. *Algorithm 4 computes the optimal solution of (P_2) with k linear inequality constraints in $O(k^2)$.*

Algorithm 4 Compute the solution $\mathbf{c}_\star$ of (P_2)

input $\mathbf{Q} \in \mathbb{S}^2$, $\mathbf{R} \in \mathbb{R}^{k\times 2}$
output $\mathbf{c}_\star = \arg\max_{\mathbf{Rc}\geq 0, \|\mathbf{c}\|_2=1} \mathbf{c}^T\mathbf{Q}\mathbf{c}$

$\mathbf{u}_1 \leftarrow$ leading eigenvector of $\mathbf{Q}$
if $\pm\mathbf{R}\mathbf{u}_1 \geq 0$ **then**
  $\mathbf{c}_\star \leftarrow \pm\mathbf{u}_1$
else
  $\mathcal{C} = \{\}$
  for $i = 1$ **to** k **do**
    $\mathbf{c}_i \leftarrow [-R_{i,2},\ R_{i,1}]^T / \|\mathbf{R}_{i,:}\|_2$
    if $\mathbf{R}(\pm\mathbf{c}_i) \geq 0$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{\pm\mathbf{c}_i\}$
    end if
  end for $\{\mathcal{C} = \emptyset \Rightarrow (P_2)$ infeasible$\}$
  $\mathbf{c}_\star \leftarrow \arg\max_{\mathbf{c}\in\mathcal{C}} \mathbf{c}^T\mathbf{Q}\mathbf{c}$
end if

Proof. There exist at most $2k+2$ candidate solutions, including $\pm\mathbf{u}_1$. Each candidate is computed in $O(1)$, and its feasibility is checked in $O(k)$. In total, the collection $\mathcal{C}$ of feasible candidate solutions is constructed in $O(k^2)$. The optimal solution is determined via exhaustive comparison among the candidates in $\mathcal{C}$ in $O(k)$. □

The arbitrary d case. We demonstrate an algorithm to solve (P_d) for any arbitrary d. Our algorithm relies on generalizing the observations and ideas of the $d = 2$ case. In particular, assuming that the feasible region is non-empty, we will show the following claim:

Claim 1. *Let $\mathbf{u}_1 \in \mathbb{R}^d$ be the leading eigenvector of $\mathbf{Q}$. If $\pm\mathbf{u}_1$ is feasible, $\mathbf{c}_\star = \pm\mathbf{u}_1$ is the optimal solution of (P_d). Otherwise, at least one of the k linear constraints holds with equality at $\mathbf{c}_\star$, i.e., $\exists i \in [k]$ such that $\mathbf{R}_{i,:}\mathbf{c}_\star = 0$.*

Proof. Let $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ be the eigenvalue decomposition of $\mathbf{Q}$: the diagonal entries of $\mathbf{\Lambda}$ coincide with the real eigenvalues of $\mathbf{Q}$, $\lambda_1, \ldots, \lambda_d$ in decreasing order, and the columns of $\mathbf{U}$ with the corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_d$. The latter form an orthonormal basis for $\mathbb{R}^d$.

Clearly, if either of $\pm\mathbf{u}_1$ is feasible, the quadratic objective attains its maximum value at $\mathbf{c}_\star = \pm\mathbf{u}_1$. In the sequel, we are concerned with the case where both $\pm\mathbf{u}_1$ are infeasible.

Consider a feasible point $\mathbf{c}_0$, $\|\mathbf{c}_0\| = 1$, such that $\mathbf{R}\mathbf{c}_0 > \mathbf{0}$. That is, $\mathbf{c}_0$ satisfies all linear constraints with strict inequality. If no such a point exists, the claim holds trivially. We will show that $\mathbf{c}_0$ cannot be optimal.

In terms of the eigenbasis, we have $\mathbf{c}_0 = \mathbf{U}\boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbf{U}^T\mathbf{c}_0 \in \mathbb{R}^d$, $\|\boldsymbol{\mu}\| = 1$. Let $\widehat{\mathbf{c}}_0$ denote the orthogonal projection of $\mathbf{c}_0$ on the subspace spanned by the trailing eigenvectors $\mathbf{u}_2, \ldots, \mathbf{u}_d$, normalized to unit length. That is, $\widehat{\mathbf{c}}_0 = \frac{1}{1-\mu_1^2}\sum_{i=2}^d \mathbf{u}_i\mu_i$. The unit norm vectors in the 2-dimensional span of $\mathbf{u}_1$ and $\widehat{\mathbf{c}}_0$ are all points of the form

$$\boldsymbol{\alpha}(\phi) = \mathbf{U}\begin{bmatrix}1 & 0\\ \mathbf{0} & \boldsymbol{\mu}_{2:d}/(1-\mu_1^2)\end{bmatrix}\begin{bmatrix}\cos(\phi)\\ \sin(\phi)\end{bmatrix},$$

for $\phi \in [0, 2\pi)$. Note that $\boldsymbol{\alpha}(0) = \mathbf{u}_1$ and $\boldsymbol{\alpha}(\pi) = -\mathbf{u}_1$ are by assumption infeasible. Therefore, points $\alpha(\phi)$ are feasible only for ϕ restricted to some interval $[\phi_1,\ \phi_2]$, with $0 \leq |\phi_1 - \phi_2| \leq \pi$. At the endpoints ϕ_1 and ϕ_2, at least one of the inequality constraints becomes active. Further, there exists a point $\phi_0 = \arccos\left(\mathbf{u}_1^T\mathbf{c}_0\right) \in [\phi_1,\ \phi_2]$, such that $\boldsymbol{\alpha}(\phi_0) = \mathbf{c}_0$. By assumption, $\mathbf{R}_{i,:}\boldsymbol{\alpha}(\phi_0) > 0, \forall i \in [k]$.

Let $Q(\phi)$ denote the objective function of (P_d) over the unit norm vectors $\alpha(\phi)$, as a function of ϕ. We will show that $Q(\phi_0) \leq \max\{Q(\phi_1), Q(\phi_2)\}$. We have

$$\begin{aligned} Q(\phi) &= \boldsymbol{\alpha}(\phi)^T\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\boldsymbol{\alpha}(\phi)\\ &= \lambda_1\cdot\cos(\phi)^2 + \frac{1}{1-\mu_1^2}\sum_{i=2}^d \mu_i^2\lambda_i\cdot\sin(\phi)^2\\ &= \lambda_1 + \frac{1}{1-\mu_1^2}\left(\boldsymbol{\mu}^T\mathbf{\Lambda}\boldsymbol{\mu} - \lambda_1\right)\sin(\phi)^2. \end{aligned}$$

Taking into account that $\lambda_1 \geq \boldsymbol{\mu}^T\mathbf{\Lambda}\boldsymbol{\mu}$, it is straightforward to verify through the first derivative w.r.t. ϕ that $Q(\phi)$ has four critical points at $\phi = 0,\ \pi/2,\ \pi$ and $3\pi/2$. One of the following holds: *(i)* $Q(\phi)$ is monotonically decreasing in $[\phi_1, \phi_2]$, *(ii)* $Q(\phi)$ is monotonically increasing in $[\phi_1, \phi_2]$, or *(iii)* a unique local minimum lies in (ϕ_1, ϕ_2). In either case, the maximum value of $Q(\phi)$ over $[\phi_1, \phi_2]$ is achieved at one of ϕ_1 and ϕ_2, which completes the proof. □

According to Claim 1, at least one linear inequality constraint holds with equality at the optimal point $\mathbf{c}_\star$. Assume that the i^{th} linear constraint is such a constraint, *i.e.*, $\mathbf{R}_{i,:}\mathbf{c}_\star = 0$. In the sequel, we investigate how this extra assumption simplifies solving (P_d).

The constraint $\mathbf{R}_{i,:}\mathbf{c} = 0$ enforces a linear dependence on the entries of $\mathbf{c}$. Let $j \in [d]$ be the index of a nonzero entry[4] of $\mathbf{R}_{i,:}$. Let $\mathbf{c}_{\setminus j} \in \mathbb{R}^{d-1}$ and $\mathbf{R}_{i,\setminus j} \in \mathbb{R}^{1\times d-1}$ denote the vectors obtained excluding the j^{th} entry of $\mathbf{c}$ and $\mathbf{R}_{i,:}$, respectively. Then,

$$\mathbf{c} = \mathbf{H}\mathbf{c}_{\setminus j}, \tag{20}$$

where

$$\mathbf{H} = \begin{bmatrix}\mathbf{I}_{j-1\times j-1} & \mathbf{0}_{j-1\times d-j}\\ -R_{i,j}^{-1}\mathbf{R}_{i,\setminus j} & \\ \mathbf{0}_{d-j\times j-1} & \mathbf{I}_{d-j\times d-j}\end{bmatrix} \in \mathbb{R}^{d\times d-1}.$$

Let $\mathbf{H} = \mathbf{U}_H\mathbf{\Sigma}_H\mathbf{V}_H^T$ be the compact singular value decomposition of the rank-$(d-1)$ matrix $\mathbf{H}$: $\mathbf{U}_H \in \mathbb{R}^{d\times d-1}$

[4] If no such j exists, the i^{th} row of $\mathbf{R}$ is the zero vector. In that case, the i^{th} linear constraint is redundant and can be omitted.

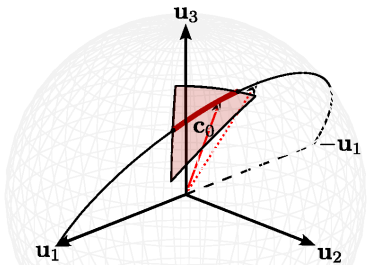


Figure 8. An instance of (P_3). Feasible solutions lie in the intersection of half-spaces with the unit sphere (highlighted region). The leading eigenvector $\pm\mathbf{u}_1$ of $\mathbf{Q}$ is not feasible. Consider any solution $\mathbf{c}_0$ in the interior of the feasible region. The highlited arc corresponds to unit length points in the span of $\mathbf{c}_0$ and $\mathbf{u}_1$. The objective value at $\mathbf{c}_0$ cannot exceed the value at the endpoints of that arc.

consists of the $d-1$ leading left singular vectors of $\mathbf{H}$, $\mathbf{V}_H \in \mathbb{R}^{d-1\times d-1}$ is a unitary matrix comprising the right singular vectors, and $\mathbf{\Sigma}_H$ is a diagonal matrix containing the $d-1$ nonzero singular values of $\mathbf{H}$.

Define $\widehat{\mathbf{c}} = \mathbf{\Sigma}_H \mathbf{V}_H^T \mathbf{c}_{\setminus j} \in \mathbb{R}^{d-1}$. Through (20), the original variable $\mathbf{c}$ can be expressed in terms of $\widehat{\mathbf{c}}$: $\mathbf{c} = \mathbf{U}_H \widehat{\mathbf{c}}$. Substituting $\mathbf{c}$ in (P_d) accordingly, we conclude that in order to compute the solution to (P_d), it suffices to compute

$$\widehat{\mathbf{c}}^{(i)} = \arg \max_{\substack{\mathbf{R}\mathbf{U}_H\widehat{\mathbf{c}}\geq \mathbf{0} \\ \|\mathbf{U}_H\widehat{\mathbf{c}}\|=1}} \widehat{\mathbf{c}}^T \left(\mathbf{U}_H^T \mathbf{Q} \mathbf{U}_H\right) \widehat{\mathbf{c}}. \tag{21}$$

The optimal solution of (P_d) will then be $\mathbf{c}^{(i)} = \mathbf{U}_H \widehat{\mathbf{c}}^{(i)}$, where the superscript is used to remind that $\mathbf{c}^{(i)}$ is optimal under the assumption that the i^{th} linear inequality constraint is active. In practice, it is not known which constraints are active at the globally optimal solution $\mathbf{c}_\star$. However, on principle, we can compute k candidates $\mathbf{c}^{(i)}$, $i \in [k]$, one for each linear inequality constraint in $\mathbf{R}$, and determine $\mathbf{c}_\star$ via exhaustive comparison.

It remains to show that we can efficiently solve (21). Due to the fact that the columns of $\mathbf{U}_H$ are orthonormal, the requirement $\|\mathbf{U}_H\widehat{\mathbf{c}}\| = 1$ is equivalent to $\|\widehat{\mathbf{c}}\| = 1$, and (21) becomes

$$\widehat{\mathbf{c}}^{(i)} = \arg \max_{\substack{\mathbf{R}^{(i)}\widehat{\mathbf{c}}\geq 0 \\ \|\widehat{\mathbf{c}}\|_2=1}} \widehat{\mathbf{c}}^T \mathbf{Q}^{(i)} \widehat{\mathbf{c}}, \qquad (P_{d-1}^{(i)})$$

with $\mathbf{R}^{(i)} = \mathbf{R}\mathbf{U}_H$ and $\mathbf{Q}^{(i)} = \mathbf{U}_H^T \mathbf{Q} \mathbf{U}_H$. One can verify that the new optimization is identical in form to (P_d), but the dimension of the unknown variable is reduced to $d-1$. Further, the i^{th} row of $\mathbf{R}^{(i)}$ is the all zero vector, effectively decreasing the number of linear constraints to $k-1$.

The algorithm. The above discussion motivates a recursion for solving (P_d). The procedure is outlined in the following steps and is formally presented in Algorithm 5.

Algorithm 5 Compute the solution $\mathbf{c}_\star$ of (P_d)

input $\mathbf{Q} \in \mathbb{S}^d$, $\mathbf{R} \in \mathbb{R}^{k\times d}$
output $\mathbf{c}_\star \in \mathbb{R}^d : \mathbf{c}_\star = \arg\max_{\mathbf{R}\mathbf{c}\geq 0, \|\mathbf{c}\|_2=1} \mathbf{c}^T\mathbf{Q}\mathbf{c}$

if $d = 2$ **then**
  $\mathbf{c}_\star \leftarrow$ solve $(P_2\,[\mathbf{Q}, \mathbf{R}])$
end if
$\mathbf{u}_1 \leftarrow$ leading eigenvector of $\mathbf{Q}$
if $\mathbf{R}\,(\pm\mathbf{u}_1) \geq 0$ **then**
  $\mathbf{c}_\star \leftarrow \pm\mathbf{u}_1$
else
  $\mathcal{C} = \{\}$ {Set of candidate solutions}
  for $i = 1$ **to** k **do**
    $j = \min\{l \in [d] : R_{il} \neq 0\}$
    $\mathbf{H} \leftarrow \begin{bmatrix} \mathbf{I}_{j-1\times j-1} & \mathbf{0}_{j-1\times d-j} \\ -R_{i,j}^{-1}\mathbf{R}_{i,\setminus j} & \\ \mathbf{0}_{d-j\times j-1} & \mathbf{I}_{d-j\times d-j} \end{bmatrix}$
    $\mathbf{U}_H, \mathbf{\Sigma}_H, \mathbf{V}_H \leftarrow$ svd$(\mathbf{H})$
    $\mathbf{Q}^{(i)} \leftarrow {\mathbf{U}_H}^T \mathbf{Q} \mathbf{U}_H$ $\{\in \mathbb{S}^{d-1}\}$
    $\mathbf{R}^{(i)} \leftarrow \mathbf{R}_{\setminus i,:}\mathbf{U}_H$ $\{\in \mathbb{R}^{k-1\times d-1}\}$
    $\widehat{\mathbf{c}}^{(i)} \leftarrow$ solve $\left(P_{d-1}^{(i)}\left[\mathbf{Q}^{(i)},\ \mathbf{R}^{(i)}\right]\right)$ $\{\in \mathbb{R}^{d-1}\}$
    $\mathbf{c}^{(i)} \leftarrow \mathbf{U}_H\widehat{\mathbf{c}}^{(i)}$ $\{\in \mathbb{R}^d\}$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{\mathbf{c}^{(i)}\}$ $\{|\mathcal{C}| \leq k\}$
  end for
  $\mathbf{c}_\star \leftarrow \arg\max_{\mathbf{c}\in\mathcal{C}} \left(\mathbf{c}^T\mathbf{Q}\mathbf{c}\right)$
end if

1. If $d = 2$, compute and return the optimal solution according Algorithm 4.
2. Compute $\mathbf{u}_1 \in \mathbb{R}^d$, the leading eigenvector of $\mathbf{Q}$. If $\pm\mathbf{u}_1$ is feasible, return $\mathbf{c}_\star = \pm\mathbf{u}_1$.
3. Otherwise, for $i = 1, \ldots, k$:
   - Form the $(d-1)$-dimensional problem $(P_{d-1}^{(i)})$, setting the i^{th} linear inequality constraint active.
   - Solve $(P_{d-1}^{(i)})$ recursively and obtain a candidate solution $\mathbf{c}^{(i)}$ of (P_d). Include $\mathbf{c}^{(i)}$ in $\mathcal{C}$, the collection of candidate solutions.
4. Return $\mathbf{c}_\star = \arg\max_{c\in\mathcal{C}} \mathbf{c}^T\mathbf{Q}\mathbf{c}$.

Lemma C.7. *Algorithm 5 solves (P_d) in $O(k^d)$.*

Proof. Let $C(d,k)$ denote the complexity of solving (P_d) with k linear inequality constraints using Algorithm 5. The leading eigenvector of $\mathbf{Q}$ is computed in $O(d^3)$ and its feasibility can be verified in $O(dk)$. Each of the k subproblems $(P_{d-1}^{(i)})$ of dimension $d-1$ with $k-1$ inequalities can be formulated in $O(d^3k)$ and solved recursively in $C(d-1,k-1)$. The maximum recursion depth is $d-2$ and the base problem (P_2) is solved in $O(k^2)$ by Alg. 4. In total, $C(d,k) = k \cdot C\,(d-1,k-1) + O(d^3k^2)$, which in turn yields $C(d,k) = O\left(k^d\right)$. □

D. Near-Linear Time Nonnegative SPCA

Alg. 1 approximates the nonnegative, k-sparse principal component of an $n \times n$ PSD matrix $\mathbf{A}$,

$$\mathbf{x}_\star = \arg\max_{\mathbf{x}\in\mathbb{S}_k^n} \quad \mathbf{x}^T\mathbf{A}\mathbf{x},$$

by efficiently solving the nonnegative sparse PCA problem on $\mathbf{A}_d$, the best rank-d approximation of $\mathbf{A}$. More precisely, Alg. 1 computes and outputs

$$\mathbf{x}_d = \arg\max_{\mathbb{S}_k^n} \mathbf{x}^T\mathbf{A}_d\mathbf{x}, \tag{22}$$

in time polynomial in n, for any constant d. The output $\mathbf{x}_d$ is a surrogate for the desired vector $\mathbf{x}_\star$.

Albeit polynomial in n, the computational complexity of Alg. 1 can be impractical even for moderate values of n. In this section, we develop Algorithm 3, a simple randomized procedure for approximating the nonnegative, k-sparse principal component of a PSD matrix in time almost linear in n. Alg. 3 relies on the same core ideas as Alg. 1: solve the nonnegative sparse PCA problem on a rank-d matrix $\mathbf{A}_d$ recasting the maximization in (22) into a series of simpler problems. But instead of computing the exact solution $\mathbf{x}_d$ of the rank-d nonnegative PCA problem, Alg. 3 settles for an approximate solution $\widehat{\mathbf{x}}_d$ computed in near-linear time. This second level of approximation introduces an additional error: $\widehat{\mathbf{x}}_d$ may be a slightly worse approximation of $\mathbf{x}_\star$ compared to $\mathbf{x}_d$. That extra approximation error, however, can be made arbitrarily small.

Lemma D.8. *Let $\mathbf{A}$ be an $n \times n$ PSD matrix given as input to Alg. 3, along with sparsity parameter $k \in [n]$ and accuracy parameters $d \in [n]$ and $\epsilon \in (0, 1]$. Let $\mathbf{A}_d$ be the best rank-d approximation of $\mathbf{A}$, and $\mathbf{x}_d$ its nonnegative, k-sparse principal component. Alg. 3 outputs a nonnegative, k-sparse, unit norm vector $\widehat{\mathbf{x}}_d$ such that*

$$\widehat{\mathbf{x}}_d^T\mathbf{A}_d\widehat{\mathbf{x}}_d \;\geq\; (1-\epsilon)\cdot {\mathbf{x}_d}^T\mathbf{A}_d\mathbf{x}_d,$$

with probability at least $1 - 1/n$, in time $O\big(\epsilon^{-d}\cdot n\log n\big)$ plus the time required to compute the d leading eigenvectors of $\mathbf{A}$.

The lemma follows from the analysis of Alg. 3, which is the focus of Section D.1, and its proof is deferred until the end of that section. According to the lemma, the output $\widehat{\mathbf{x}}_d$ of Alg. 3 is a factor $(1-\epsilon)$ approximation of $\mathbf{x}_d$, the nonnegative, k-sparse principal component of $\mathbf{A}_d$, in terms of explained variance on the rank-d matrix $\mathbf{A}_d$. Our ultimate goal, however, is to characterize the quality of $\widehat{\mathbf{x}}_d$ as a surrogate for $\mathbf{x}_\star$, the nonnegative, k-sparse principal component of $\mathbf{A}$. The approximation guarantees of Alg. 3 are established in the following theorem.

Theorem 2. *For any $n\times n$ PSD matrix $\mathbf{A}$, sparsity parameter k, and accuracy parameters $d \in [n]$ and $\epsilon \in (0, 1]$, Alg. 3 outputs a nonnegative, k-sparse, unit norm vector $\widehat{\mathbf{x}}_d$ such that*

$$\widehat{\mathbf{x}}_d^T\mathbf{A}\widehat{\mathbf{x}}_d \;\geq\; (1-\epsilon)\cdot\rho_d\cdot {\mathbf{x}_\star}^T\mathbf{A}\mathbf{x}_\star,$$

with probability at least $1 - 1/n$, in time $O\big(\epsilon^{-d}\cdot n\log n\big)$ plus the time required to compute the d leading eigenvectors of $\mathbf{A}$.

Proof. For the output $\widehat{\mathbf{x}}_d$ of Alg. 3, we have

$$\begin{aligned}
\widehat{\mathbf{x}}_d^T\mathbf{A}\widehat{\mathbf{x}}_d \;&=\; \widehat{\mathbf{x}}_d^T\mathbf{A}_d\widehat{\mathbf{x}}_d &&+\, \widehat{\mathbf{x}}_d^T\left(\mathbf{A}-\mathbf{A}_d\right)\widehat{\mathbf{x}}_d\\
&\overset{(\alpha)}{\geq}\; (1-\epsilon)\cdot {\mathbf{x}_d}^T\mathbf{A}_d\mathbf{x}_d &&+\, \widehat{\mathbf{x}}_d^T\left(\mathbf{A}-\mathbf{A}_d\right)\widehat{\mathbf{x}}_d\\
&\overset{(\beta)}{\geq}\; (1-\epsilon)\cdot {\mathbf{x}_d}^T\mathbf{A}_d\mathbf{x}_d\\
&=\; (1-\epsilon)\cdot \mathsf{OPT}_d\\
&=\; (1-\epsilon)\cdot\rho_d\cdot\mathsf{OPT},
\end{aligned}$$

where inequality (α) follows from Lemma D.8, and (β) from the fact that $\mathbf{A}-\mathbf{A}_d$ is a PSD matrix. Note that by Lemma A.5, $\rho_d = \mathsf{OPT}_d/\mathsf{OPT}$ satisfies

$$\rho_d \;\geq\; \max\left\{\frac{k}{2n},\; \frac{1}{1+2\frac{n}{k}\lambda_{d+1}/\lambda_1}\right\}.$$

The complexity of Alg. 3 is established in Lemma D.8, which completes the proof. □

D.1. Analysis of Algorithm 3

In this subsection, we examine Alg. 3 in detail and gradually build towards establishing Lemma D.8.

Given an $n \times n$ PSD matrix $\mathbf{A}$ and an accuracy parameter d, Alg. 3 first computes the d leading eigenvectors of $\mathbf{A}$ to obtain the rank-d approximation $\mathbf{A}_d$. Let $\mathbf{V}$ be an $n \times d$ square root of $\mathbf{A}_d$. That is, $\mathbf{A}_d = \mathbf{V}\mathbf{V}^T$. In subsection 3.2, we showed that the rank-d nonnegative sparse PCA problem on $\mathbf{A}_d$ can be written as

$$\max_{\mathbf{x}\in\mathbb{S}_k^n}\mathbf{x}^T\mathbf{A}_d\mathbf{x} = \max_{\mathbf{c}\in\mathbb{S}^d,\,\mathbf{x}\in\mathbb{S}_k^n}\Big(\left(\mathbf{V}\mathbf{c}\right)^T\mathbf{x}\Big)^2. \tag{23}$$

For a fixed $\mathbf{c}$, the optimal $\mathbf{x}$ can be easily determined as described in Section 3.1. In principle, scanning all vectors $\mathbf{c}$ on the surface of the d-dimensional unit sphere $\mathbb{S}^d$ would suffice to detect the nonnegative, k-sparse principal component $\mathbf{x}_d$.

Alg. 3 approximately solves the double maximization in (23) considering a finite set of $m = O(\epsilon^{-d}\cdot\log n)$ points $\mathbf{c}_1,\ldots,\mathbf{c}_m$ drawn randomly and independently, uniformly distributed over $\mathbb{S}^d$. Each random point $\mathbf{c}_i$ corresponds to an n-dimensional vector $\mathbf{a}_i = \mathbf{V}\mathbf{c}_i$ in the range of $\mathbf{A}_d$, for which Alg. 3 solves the rank-1 nonnegative sparse PCA

problem

$$\max_{\mathbf{x}\in\mathbb{S}^n_k}\left((\mathbf{V}\mathbf{c}_i)^T\mathbf{x}\right)^2 = \max_{\mathbf{x}\in\mathbb{S}^n_k}\left(\mathbf{a}_i{}^T\mathbf{x}\right)^2.$$

The rank-1 problem is solved in time $O(n)$ as described in Section 3.1, and yields a candidate solution $\mathbf{x}$. Alg. 3 outputs the candidate that maximizes $\|\mathbf{V}^T\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{A}_d\mathbf{x}$.

In the following, we argue that the $m = O(\epsilon^{-d}\cdot\log n)$ random samples suffice to establish the approximation guarantees of Lemma D.8, and in turn Theorem 2.

Randomized ϵ-nets. An ϵ-net on the unit sphere $\mathbb{S}^d$ is a set $\mathcal{N}^d_\epsilon$ of points on $\mathbb{S}^d$ such that for any point on $\mathbb{S}^d$ there exists a point in $\mathcal{N}^d_\epsilon$ within euclidean distance ϵ. More formally,

Def. 2. *An ϵ-net of $\mathbb{S}^d$ is a set $\mathcal{N}^d_\epsilon \subset \mathbb{S}^d$ such that*

$$\forall\mathbf{c}\in\mathbb{S}^d,\ \ \exists\ \widehat{\mathbf{c}}\in\mathcal{N}^d_\epsilon : \|\mathbf{c}-\widehat{\mathbf{c}}\|_2 \le \epsilon.$$

Consider a $(\epsilon/2)$-net $\mathcal{N}^d_{\epsilon/2}$ on $\mathbb{S}^d$ for some given constant $0 < \epsilon \le 1$. The following lemma states that if we solve the maximization in (23) over the points $\mathbf{c}$ in the finite set of points $\mathcal{N}^d_{\epsilon/2}$ instead of the entire sphere $\mathbb{S}^d$, we obtain a solution that is within a factor $(1-\epsilon)$ from OPT_d.

Lemma D.9. *Let $\mathcal{N}^d_{\epsilon/2}$ be a $\epsilon/2$-net of $\mathbb{S}^d$. Then,*

$$(1-\epsilon)\cdot\mathsf{OPT}_d \ \le\ \max_{\mathbf{c}\in\mathcal{N}^d_{\epsilon/2}}\max_{\mathbf{x}\in\mathbb{S}^n_k}\left(\mathbf{c}^T\mathbf{V}^T\mathbf{x}\right)^2 \ \le\ \mathsf{OPT}_d.$$

Proof. The upper bound follows from the fact that $\mathcal{N}^d_{\epsilon/2}\subseteq\mathbb{S}^d$. For the lower bound, let $(\mathbf{x}_d,\,\mathbf{c}_d)$ denote the optimal solution of (23), *i.e.*,

$$\mathsf{OPT}_d = (\mathbf{c}_d\mathbf{V}^T\mathbf{x}_d)^2.$$

By definition, the $\epsilon/2$-net $\mathcal{N}^d_{\epsilon/2}$ contains a vector $\widehat{\mathbf{c}}_d$ such that $\mathbf{c}_d = \widehat{\mathbf{c}}_d + \mathbf{r}$ for some $\mathbf{r}\in\mathbb{R}^d$ with $\|\mathbf{r}\|\le\epsilon/2$. Then,

$$\begin{aligned}\sqrt{\mathsf{OPT}_d} &= \mathbf{c}_d^T\mathbf{V}^T\mathbf{x}_d = (\widehat{\mathbf{c}}_d+\mathbf{r})^T\mathbf{V}^T\mathbf{x}_d\\ &\overset{(\alpha)}{\le} \widehat{\mathbf{c}}_d^T\mathbf{V}^T\mathbf{x}_d + \frac{\epsilon}{2}\cdot\|\mathbf{V}^T\mathbf{x}_d\|\\ &= \widehat{\mathbf{c}}_d^T\mathbf{V}^T\mathbf{x}_d + \frac{\epsilon}{2}\cdot\sqrt{\mathsf{OPT}_d},\end{aligned}\tag{24}$$

where (α) is due to the triangle inequality, the Cauchy-Schwartz inequality, and the fact that $\|\mathbf{r}\|\le\epsilon/2$. From (24), it follows that

$$\left(\widehat{\mathbf{c}}_d^T\mathbf{V}^T\mathbf{x}_d\right)^2 \ge (1-\epsilon/2)^2\cdot\mathsf{OPT}_d \ge (1-\epsilon)\cdot\mathsf{OPT}_d.$$

Noting that

$$\max_{\mathbf{c}\in\mathcal{N}^d_{\epsilon/2}}\max_{\mathbf{x}\in\mathbb{S}^n_k}\left(\mathbf{c}^T\mathbf{V}^T\mathbf{x}\right)^2 \ \ge\ \left(\widehat{\mathbf{c}}_d^T\mathbf{V}^T\mathbf{x}_d\right)^2$$

completes the proof. □

There are many constructions for ϵ-nets on the sphere, both deterministic and randomized (Rogers, 1957; Böröczky Jr & Wintsche, 2003; Dumer, 2007). In the following we review a simple randomized construction, initially studied by Wyner (Wyner, 1967) in the asymptotic $d\to\infty$ regime. First, note the following existential result.

Lemma D.10 ((Vershynin, 2010))**.** *For any $0<\epsilon\le 1$, there exists an ϵ-net $\mathcal{N}^d_\epsilon$ of the unit sphere $\mathbb{S}^d$ with cardinality at most $m_{\epsilon,d}\le(1+2/\epsilon)^d$.*

Consider a set of sphere-caps of radius $\widehat{\epsilon}$ centered at the points of the $\widehat{\epsilon}$-net $\mathcal{N}^d_{\widehat{\epsilon}}$. The caps cover the entire sphere surface. It can be easily shown, based on a simple triangle inequality, that an arbitrary collection of points comprising at least one point from each cap, forms a $(2\widehat{\epsilon})$-net. Further, using standard balls and bins arguments, we conclude that randomly and independently drawing $O(m_{\widehat{\epsilon},d}\cdot\log(n\cdot m_{\widehat{\epsilon},d}))$ points uniformly distributed over $\mathbb{S}^d$ suffice for at least one random point to lie in each sphere cap with probability at least $1-1/n$. That is, $O\left(\widehat{\epsilon}^{-d}\cdot\log n\right)$ points suffice to form a $(2\widehat{\epsilon})$-net. Hence, for $\widehat{\epsilon}=\epsilon/4$, we will obtain an $\epsilon/2$-net.

Lemma D.11. *Randomly and independently drawing $O\left(\epsilon^{-d}\cdot\log n\right)$ points uniformly distributed on $\mathbb{S}^d$ suffices to form an $\epsilon/2$-net of $\mathbb{S}^d$, with probability at least $1-1/n$.*

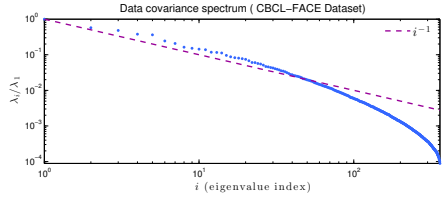
Proof of Lemma D.8 By Lemma D.11, the $O(\epsilon^{-d}\cdot\log n)$ points drawn randomly and independently uniformly over $\mathbb{S}^d$, form an $\epsilon/2$-net $\mathcal{N}^d_{\epsilon/2}$, with probability at least $1-1/n$. Alg. 3 solves the double maximization problem in (23) over the points in $\mathcal{N}^d_{\epsilon/2}$, and outputs a nonnegative, k-sparse, unit-norm vector $\widehat{\mathbf{x}}_d$. By Lemma D.9, $\widehat{\mathbf{x}}_d$ is within factor of $(1-\epsilon)$ from OPT_d, which proves the desired approximation guarantee.

Alg. 3 examines $O(\epsilon^{-d}\log n)$ points in the range of the $n\times d$ matrix $\mathbf{V}$. Each sample yields a candidate solution computed in $O(n)$. The total computational complexity is $O(\epsilon^{-d}\cdot n\log n)$, plus the time required to compute the d columns of $\mathbf{V}$, *i.e.*, the d leading eigenvectors of $\mathbf{A}$, which completes the proof. ■

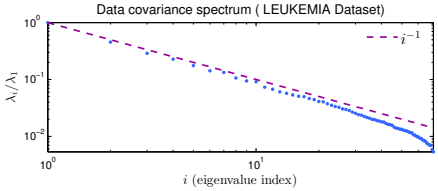
E. Power Law Spectral Decay

The approximation guarantees of our algorithm are contingent on the spectrum of the data covariance matrix: the sharper the eigenvalue decay, the tighter the approximation. In this section, we provide empirical evidence that real datasets often exhibit a steep decline in the spectrum of their empirical covariance matrix. In particular, the eigenvalues of the later approximately decay according to a power law:

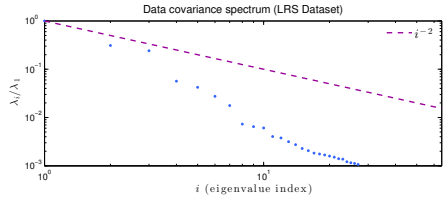
$$\lambda_i \le c\cdot\lambda_1\cdot i^\alpha,$$



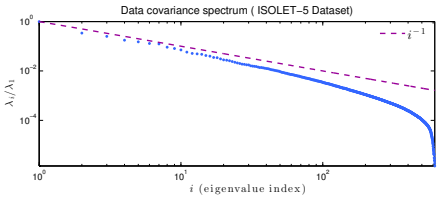
(a) CBCL Face Dataset



(b) Leukemia Dataset



(c) Low Res. Spectr. Dataset



(d) Isolet Dataset

Figure 9. Spectrum of the empirical covariance matrix of various datasets. The eigenvalues exhibit approximately power law decay. (Datasets 9(b), 9(c) and 9(d) are available at (Bache & Lichman, 2013)).

for some constant c and $\alpha \geq 1$.

Fig. 9 depicts the leading eigenvalues of the empirical covariance matrix of various datasets, normalized by the maximum eigenvalue, λ_1. In all depicted cases, the eigenvalues can be upper bounded by a power law decay function.

F. NP-Hardness of Nonnegative PCA

We provide a proof of the NP-hardness of the nonnegative (and in turn the nonnegative sparse) PCA problem with a reduction from the problem of checking whether a matrix is copositive. A matrix $\mathbf{M} \in \mathbb{S}^n$ is copositive iff $\mathbf{x}^T\mathbf{M}\mathbf{x} \geq 0$ for all vectors $\mathbf{x}$ in the nonnegative orthant:

$$\mathbf{M} \text{ is copositive } \Leftrightarrow \ \mathbf{M} \in \mathbb{S}^n : \mathbf{x}^T\mathbf{M}\mathbf{x} \geq 0, \ \forall \mathbf{x} \geq 0.$$

Checking whether a matrix is copositive is a co-NP complete (Murty & Kabadi, 1987) decision problem: any vector $\mathbf{x}$ for which $\mathbf{x}^T\mathbf{M}\mathbf{x} < 0$ serves as a certificate to verify in polynomial time that $\mathbf{M}$ is *not* copositive.

In order to check whether a matrix $\mathbf{M}$ is copositive, it suffices to minimize the quadratic $\mathbf{x}^T\mathbf{M}\mathbf{x}$ over all nonnegative vectors $\mathbf{x}$: $\mathbf{M}$ is *not* copositive if and only if the minimum value is negative. Observing that the sign of the quadratic form does not change if $\mathbf{x}$ is scaled by a positive scalar, without loss of generality restrict our attention to unit norm vectors $\mathbf{x}$ and solve

$$\mathsf{OPT} \ = \ \min_{\substack{\mathbf{x} \geq \mathbf{0} \\ \|\mathbf{x}\|=1}} \mathbf{x}^T\mathbf{M}\,\mathbf{x}. \qquad (P)$$

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\mathbf{M}$ in decreasing order. The matrix $\overline{\mathbf{M}} = \lambda_1\mathbf{I} - \mathbf{M}$ is positive semidefinite: its eigenvalues are $\lambda_1 - \lambda_i \geq 0$, for $1 \leq i \leq n$. Moreover, for any unit length vector $\mathbf{x}$, $\mathbf{x}^T\overline{\mathbf{M}}\mathbf{x} = \lambda_1 - \mathbf{x}^T\mathbf{M}\mathbf{x}$. Hence, (P) is equivalent to

$$\overline{\mathsf{OPT}} = \max_{\substack{\mathbf{x} \geq 0 \\ \|\mathbf{x}\|=1}} \mathbf{x}^T\overline{\mathbf{M}}\mathbf{x}, \qquad (\overline{P})$$

the nonnegative PCA problem on the PSD matrix $\overline{\mathbf{M}}$. Solving $(\overline{P})$ suffices to check whether $\mathbf{M}$ is copositive since $\mathsf{OPT} \leq 0$ if and only if $\overline{\mathsf{OPT}} \geq \lambda_1$. We conclude that $(\overline{P})$ is NP-hard.

Appendix References

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Böröczky Jr, Károly and Wintsche, Gergely. Covering the sphere by equal spherical balls. In *Discrete and Computational Geometry*, pp. 235–251. Springer, 2003.

Dumer, Ilya. Covering spheres with spheres. *Discrete & Computational Geometry*, 38(4):665–679, 2007.

Murty, Katta G. and Kabadi, Santosh N. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

Rogers, CA. A note on coverings. *Mathematika*, 4(01):1–6, 1957.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wyner, Aaron D. Random packings and coverings of the unit n-sphere. *Bell System Technical Journal*, 46(9):2111–2118, 1967.