Sparse PCA from Sparse Linear Regression

Guy Bresler
MIT
guy@mit.edu

Sung Min Park
MIT
sp765@mit.edu

Mădălina Persu
Two Sigma,[*] MIT
mpersu@mit.edu

Abstract

Sparse Principal Component Analysis (SPCA) and Sparse Linear Regression (SLR) have a wide range of applications and have attracted a tremendous amount of attention in the last two decades as canonical examples of statistical problems in high dimension. A variety of algorithms have been proposed for both SPCA and SLR, but an explicit connection between the two had not been made. We show how to efficiently transform a black-box solver for SLR into an algorithm for SPCA: assuming the SLR solver satisfies prediction error guarantees achieved by existing efficient algorithms such as those based on the Lasso, the SPCA algorithm derived from it achieves near state of the art guarantees for testing and for support recovery for the single spiked covariance model as obtained by the current best polynomial-time algorithms. Our reduction not only highlights the inherent similarity between the two problems, but also, from a practical standpoint, allows one to obtain a collection of algorithms for SPCA directly from known algorithms for SLR. We provide experimental results on simulated data comparing our proposed framework to other algorithms for SPCA.

1 Introduction

Principal component analysis (PCA) is a fundamental technique for dimension reduction used widely in data analysis. PCA projects data along a few directions that explain most of the variance of observed data. One can also view this as linearly transforming the original set of variables into a (smaller) set of uncorrelated variables called principal components.

Recent work in high-dimensional statistics has focused on sparse principal component analysis (SPCA), as ordinary PCA estimates become inconsistent in this regime [22]. In SPCA, we restrict the principal components to be sparse, meaning they have only a few nonzero entries in the original basis. This has the advantage, among others, that the components are more interpretable [23, 49], while components may no longer be uncorrelated. We study SPCA under the Gaussian *(single) spiked covariance model* introduced by [21]: we observe n samples of a random variable X distributed according to a Gaussian distribution $\mathcal{N}(0, I_d + \theta u u^\top)$, where $||u||_2 = 1$ with at most k nonzero entries,[2] I_d is the $d \times d$ identity matrix, and θ is the signal-to-noise parameter. We study two settings of the problem, hypothesis testing and support recovery.

Sparsity assumptions have played an important role in a variety of other problems in high-dimensional statistics, in particular linear regression. Linear regression is also ill-posed in high dimensions, so via imposing sparsity on the regression vector we recover tractability.

Though the literature on two problems are largely disjoint, there is a striking similarity between the two problems, in particular when we consider statistical and computational trade-offs. The

---

[2]Sometimes we will write this latter condition as $u \in B_0(k)$ where $B_0(k)$ is the "ℓ_0-ball" of radius k.

natural information-theoretically optimal algorithm for SPCA [4] involves searching over all possible supports of the hidden spike. This bears resemblance to the minimax optimal algorithm for SLR [35], which optimizes over all sparse supports of the regression vector. Both problems appear to exhibit gaps between statistically optimal algorithms and known computationally efficient algorithms, and conditioned on relatively standard complexity assumptions, these gaps seem irremovable [3, 44, 47].

1.1 Our contributions

In this paper we give algorithmic evidence that this similarity is likely not a coincidence. Specifically, we give a simple, general, and efficient procedure for transforming a black-box solver for sparse linear regression to an algorithm for SPCA. At a high level, our algorithm tries to predict each coordinate[3] linearly from the rest of the coordinates using a black-box algorithm for SLR. The advantages of such a black-box framework are two fold: theoretically, it highlights a structural connection between the two problems; practically, it allows one to simply plug in any of the vast majority of solvers available for SLR and directly get an algorithm for SPCA with provable guarantees. In particular,

- **For hypothesis testing:** we match state of the art provable guarantee for computationally efficient algorithms; our algorithm successfully distinguishes between isotropic and spiked Gaussian distributions at signal strength $\theta \gtrsim \sqrt{\frac{k^2 \log d}{n}}$. This matches the phase transition of diagonal thresholding [22] and Minimal Dual Perturbation [4] up to constant factors.

- **For support recovery:** for general p and n, when each non-zero entry of u is at least $\Omega(1/\sqrt{k})$ (a standard assumption in the literature), our algorithm succeeds with high probability for signal strength $\theta \gtrsim \sqrt{\frac{k^2 \log d}{n}}$, which is nearly optimal.[4]

- In experiments, we demonstrate that using popular existing SLR algorithms as our black-box results in reasonable performance.

- We theoretically and empirically illustrate that our SPCA algorithm is also robust to *rescaling of the data*, for instance by using a Pearson correlation matrix instead of a covariance matrix.[5] Many iterative methods rely on initialization via first running diagonal thresholding, which filters variables with higher variance; rescaling renders diganoal thresholding useless, so in some sense our framework is more robust.

2 Preliminaries

2.1 Problem formulation for SPCA

Hypothesis testing Here, we want to distinguish whether X is distributed according to an isotropic Gaussian or a spiked covariance model. That is, our null and alternate hypotheses are:

$$H_0 : X \sim \mathcal{N}(0, I_d) \text{ and } H_1 : X \sim \mathcal{N}(0, I_d + \theta uu^\top),$$

Our goal is to design a test $\psi \colon \mathbb{R}^{n \times d} \to \{0, 1\}$ that discriminates H_0 and H_1. More precisely, we say that ψ discriminates between H_0 and H_1 with probability $1 - \delta$ if both type I and II errors have a probability smaller than δ:

$$\mathbf{P}_{H_0}(\psi(X) = 1) \leq \delta \text{ and } \mathbf{P}_{H_1}(\psi(X) = 0) \leq \delta.$$

We assume the following additional condition on the spike u:

(C1) $c_{min}^2/k \leq u_i^2 \leq 1 - c_{min}^2/k$ for at least one $i \in [d]$ where $c_{min} > 0$ is some constant.

---

[3]From here on, we will use "coordinate" and "variable" interchangeably.

[4]In the scaling limit $d/n \to \alpha$ as $d, n \to \infty$, the covariance thresholding algorithm [15] theoretically succeeds at a signal strength that is an order of $\sqrt{\log d}$ smaller. However, our experimental results indicate that with an appropriate choice of black-box, our Q algorithm outperforms covariance thresholding

[5]Solving SPCA based on the correlation matrix was suggested in a few earlier works [49, 41].

The above condition says that at least one coordinate has enough mass, yet the mass is not entirely concentrated on just that singlecoordinate. Trivially, we always have at least one $i \in [d]$ s.t. $u_i^2 \geq 1/k$, but this is not enough for our regression setup, since we want at least one other coordinate j to have sufficient correlation with coordinate i. We remark that the above condition is a very mild technical condition. If it were violated, almost all of the mass of u is on a single coordinate, so a simple procedure for testing the variance (which is akin to diagonal thresholding) would suffice.

Support recovery The goal of support recovery is to identify the support of u from our samples $\mathbf{X}$. More precisely, we say that a support recovery algorithm succeeds if the recovered support $\widehat{S}$ is the same as S, the true support of u. As standard in the literature [1, 31], we need to assume a minimal bound on the size of the entries of u in the support.

For our support recovery algorithm, we will assume the following condition (note that it implies Condition (C1) and is much stronger):

(C2) $|u_i| \geq c_{min}/\sqrt{k}$ for some constant $0 < c_{min} < 1 \; \forall i \in [d]$

Though the settings are a bit different, this minimal bound along with our results are consistent with lower bounds known for sparse recovery. These lower bounds ([18, 42]; bound of [18] is a factor of k weaker) imply that the number of samples must grow roughly as $n \gtrsim (1/u_{min}^2)k \log d$ where u_{min} is the smallest entry of our signal u normalized by $1/\sqrt{k}$, which is qualitativley the same threhold required by our theorems.

2.2 Background on SLR

In linear regression, we observe a response vector $y \in \mathbb{R}^n$ and a design matrix $\mathbb{X} \in \mathbb{R}^{n \times d}$ that are linked by the linear model $y = \mathbb{X}\beta^* + w$, where $w \in \mathbb{R}^n$ is some form of observation noise, typically with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Our goal is to recover β^* given noisy observations y. While the matrices $\mathbb{X}$ we consider arise from a (random) correlated design (as analyzed in [42], [43]), it will make no difference to assume the matrices are deterministic by conditioning, as long as the distribution of the design matrix and noise are independent, which we will demonstrate in our case. Most of the relevant results on sparse linear regression pertain to deterministic design.

In sparse linear regression, we additionally assume that β^* has only k non-zero entries, where $k \ll d$. This makes the problem well posed in the high-dimensional setting. Commonly used performance measures for SLR are tailored to prediction error ($1/n\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2$ where $\widehat{\beta}$ is our guess), support recovery (recovering support of β^*), or parameter estimation (minimizing $\|\beta^* - \widehat{\beta}\|$ under some norm). We focus on prediction error, analyzed over random realizations of the noise. There is a large amount of work on SLR and we defer a more in-depth overview to Appendix A.

Most efficient methods for SLR impose certain conditions on $\mathbb{X}$. We focus on the *restricted eigenvalue* condition, which roughly stated makes the prediction loss strongly convex near the optimum:

Definition 2.1 (Restricted eigenvalue [47])**.** First define the cone $\mathbb{C}(S) = \{\beta \in \mathbb{R}^d \mid \|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1\}$, where S^c denotes the complement, β_T is β restricted to the subset T. The *restricted eigenvalue* (RE) constant of $\mathbb{X}$, denoted $\gamma(\mathbb{X})$, is defined as the largest constant $\gamma > 0$ s.t.

$$1/n\|\mathbb{X}\beta\|_2^2 \geq \gamma\|\beta\|_2^2 \quad \text{for all } \beta \in \bigcup_{|S|=k, S \subseteq [d]} \mathbb{C}(S)$$

For more discussion on the restricted eigenvalue, see Appendix A.

Black-box condition Given the known guarantees on SLR, we define a condition that is natural to require on the guarantee of our SLR black-box, which is invoked as $\mathsf{SLR}(y, \mathbb{X}, k)$.

Condition 2.2 (Black-box condition)**.** Let $\gamma(\mathbb{X})$ denote the restricted eigenvalue of $\mathbb{X}$. There are universal constants c, c', c'' such that $\mathsf{SLR}(y, \mathbb{X}, k)$ outputs $\widehat{\beta}$ that is k-sparse and satisfies:

$$\frac{1}{n}\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq \frac{c}{\gamma(\mathbb{X})^2}\frac{(\sigma^2 k \log d)}{n} \quad \forall \beta^* \in B_0(k) \text{ w.p.} \geq 1 - c'\exp(-c''k \log d)$$

3

3 Algorithms and main results

We first discuss how to view samples from the spiked covariance model in terms of a linear model. We then give some intuition motivating our statistic. Finally, we state our algorithms and main theorems, and give a high-level sketch of the proof.

3.1 The linear model

Let $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ be n i.i.d. samples from the spiked covariance model; denote as $\mathbf{X} \in \mathbb{R}^{n \times d}$ the matrix whose rows are $X^{(i)}$. Intuitively, if variable i is contained in the support of the spike, then the rest of the support should allow to provide a nontrivial prediction for $\mathbf{X}_i$ since variables in the support are correlated. Conversely, for i not in the support (or under the isotropic null hypothesis), all of the variables are independent and other variables are useless for predicting $\mathbf{X}_i$. So we regress $\mathbf{X}_i$ onto the rest of the variables.

Let $\mathbf{X}_{-i}$ denote the matrix of samples in the SPCA model with the ith column removed. For each column i, we can view our data as coming from a linear model with design matrix $\mathbb{X} = \mathbf{X}_{-i}$ and the response variable $y = \mathbf{X}_i$.

The "true" regression vector depends on i. Under the alternate hypothesis H_1, if $i \in S$, we can write $y = \mathbb{X}\beta^* + w$ where $\beta^* = \frac{\theta u_i}{1+(1-u_i^2)\theta} u_{-i}$ and $w \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1 + \frac{\theta u_i^2}{1+(1-u_i^2)\theta}$.[6] If $i \notin S$, and for any $i \in [d]$ under the null hypothesis, $y = w$ where $w = \mathbf{X}_i \sim \mathcal{N}(0, 1)$ (implicitly $\beta^* = \mathbf{0}$).

3.2 Designing the test statistic

Based on the linear model above, we want to compute a test statistic that will indicate when a coordinate i is on support. Intuitively, we predictive power of our linear model should be higher when i is on support. Indeed, a calculation shows that the variance in X_i is reduced by approximately θ^2/k. We want to measure this reduction in noise to detect when i is on support or not.

Suppose for instance that we have access to β^* rather than $\widehat{\beta}$ (note that this is not possible in practice since we do not know the support!). Since we want to measure the reduction in noise when the variable is on support, as a first step we might try the following statistic:

$$Q_i = \frac{1}{n}\|y - \mathbb{X}\beta^*\|_2^2$$

Unfortunately, this statistic will not be able to distinguish the two hypotheses, as the reduction in the above error is too small (on the order of θ^2/k compared to overall order of $1+\theta$), so deviation due to random sampling will mask any reduction in noise. We can fix this by adding the variance term $\|y\|^2$:

$$Q_i = \frac{1}{n}\|y\|_2^2 - \frac{1}{n}\|y - \mathbb{X}\beta^*\|_2^2$$

On a more intuitive level, including $\|y\|_2^2$ allows us to measure the relative *gain* in predictive power without being penalized by a possibly large variance in y. Fluctuations in y due to noise will typically be canceled out in the difference of terms in Q_i, minimizing the variance of our statistic.

We have to add one final fix to the above estimator. We obviously do not have access to β^*, so we must use the estimate $\widehat{\beta} = \mathsf{SLR}(y, \mathbb{X}, k)$ ($y, \mathbb{X}$ are as defined in Section 3.1) which we get from our black-box. As our analysis shows, this substitution does not affect much of the discriminative power of Q_i as long as the SLR black-box satisfies prediction error guarantees stated in Condition 2.2. This gives our final statistic:[7]

$$Q_i = \frac{1}{n}\|y\|_2^2 - \frac{1}{n}\|y - \mathbb{X}\widehat{\beta}\|_2^2.$$

3.3 Algorithms

Below we give algorithms for hypothesis testing and for support recovery, based on the Q statistic:

---

[6] By the theory of linear minimum mean-square-error (LMMSE) confirms that this choice of β^* minimizes the error σ^2. See Appendix B.1, B.2 for details of this calculation.

[7] As pionted out by a reviewer, Note that this statistic is actually equivalent to R^2 up to rescaling by sample variance. Note that our formula is slightly different though as we use the sample variance computed with population mean as opposed to sample mean, as the mean is known to be zero.

Algorithm 1 *Q*-hypothesis testing

Input: $\mathbf{X} \in \mathbb{R}^{d \times n}, k$
Output: $\{0, 1\}$ → dimension
for $i = 1, \ldots, d$ **do**
  $\widehat{\beta}_i = \mathsf{SLR}(\mathbf{X}_i, \mathbf{X}_{-i}, k)$
  $Q_i = \frac{1}{n}\|\mathbf{X}_i\|_2^2 - \frac{1}{n}\|\mathbf{X}_i - \mathbf{X}_{-i}\widehat{\beta}_i\|_2^2$
  if $Q_i > \frac{13k \log \frac{d}{k}}{n}$ **then**
    return 1 → 하나가 초과면 1 출력
  end if
end for
Return 0

Algorithm 2 *Q*-support recovery

Input: $\mathbf{X} \in \mathbb{R}^{d \times n}, k$
$\widehat{S} = \varnothing$
for $i = 1, \ldots, d$ **do**
  $\widehat{\beta}_i = \mathsf{SLR}(\mathbf{X}_i, \mathbf{X}_{-i}, k)$
  $Q_i = \frac{1}{n}\|\mathbf{X}_i\|_2^2 - \frac{1}{n}\|\mathbf{X}_i - \mathbf{X}_{-i}\widehat{\beta}_i\|_2^2$
  if $Q_i > \frac{13k \log \frac{d}{k}}{n}$ **then**
    $\widehat{S} \leftarrow \widehat{S} \cup \{i\}$
  end if
end for
Return $\widehat{S}$ ↳ 하나가 초과면 합친다

Below we summarize our guarantees for the above algorithms. The proofs are simple, but we defer them to Appendix C.

Theorem 3.1 (Hypothesis test). *Assume we have access to* SLR *that satisfies Condition 2.2 and with runtime $T(d, n, k)$ per instance. Under Condition (C1), there exist universal constants c_1, c_2, c_3, c_4 s.t. if $\theta^2 > \frac{c_1}{c_{min}^2} \frac{k^2 \log d}{n}$ and $n > c_2 k \log d$, Algorithm 1 outputs ψ s.t.*

$$\mathbf{P}_{H_0}(\psi(X) = 1) \vee \mathbf{P}_{H_1}(\psi(X) = 0) \leq c_3 \exp(-c_4 k \log d)$$

in time $O(dT + d^2 n)$.

Theorem 3.2 (Support recovery). *Under the same conditions as above plus Condition (C2), if $\theta^2 > \frac{c_1}{c_{min}^2} \frac{k^2 \log d}{n}$, Algorithm 2 above finds $\widehat{S} = S$ with probability at least $1 - c_3 \exp(-c_4 k \log d)$ in time $O(dT + d^2 n)$.*

3.4 Comments → 증명 가능

RE for sample design matrix Because population covariance $\Sigma = \mathbb{E}[\mathbb{X}\mathbb{X}^\top]$ has minimum eigenvalue 1, with high probability the sample design matrix $\mathbb{X}$ has constant restricted eigenvalue value given enough samples, i.e. n is large enough (see Appendix B.3 for more details), and the prediction error guarantee of Condition 2.2 will be good enough for our analysis.

Running time The runtime of both Algorithm 1 and 2 is $\tilde{O}(nd^2)$. The discussion presented at the end of Appendix C details why this is competitive for (single spiked) SPCA, at least theoretically.

Unknown sparsity Throughout the paper, we assume that the sparsity level k is known. However, if k is unknown, standard techniques could be used to adaptively find approximate values of k ([16]). For instance, for hypothesis testing, we can start with an initial overestimate k', and keep halving until we get enough coordinates i with Q_i that passes the threshold for the given k'.

↳ k 사전에 모르면

Robustness of Q statistic to rescaling

→ 정확한 scale 필요 X

Intuitively, our algorithms for detecting correlated structure in data should be invariant to rescaling of the data; the precise scale or units for which one variable is measured should not have an impact on our ability to find meaningful structure underlying the data. Our algorithms based on Q are robust to rescaling, perhaps unsurprisingly, since correlations between variables in the support remain under rescaling. 다르기

On the other hand, diagonal thresholding, an often-used preprocessing step for SPCA, which filters variables strictly based on variance, would trivially fail under rescaling. This illustrates a strength of our framework over other existing algorithms for SPCA. ↳ 강조

error

Below we show explicitly that Q statistics are indeed robust to rescaling: Let $\widetilde{X} = DX$ be the rescaling of X, where D is some diagonal matrix. Let D_S be D restricted to rows and columns in S. Note that $\widetilde{\Sigma}$, the covariance matrix of the rescaled data, is just $D\Sigma D$ by expanding the definition.

normalize

5

Similarly, note $\widetilde{\Sigma}_{2:d,1} = D_1 D_{2:d}\Sigma_{2:d,1}$ where $D_{2:d}$ denotes D without row and column 1. Now, recall the term which dominated our analysis of Q_i under H_1, $(\beta^*)^\top \Sigma_{2:d}\beta^*$, which was equal to

$$\Sigma_{1,2:d}\Sigma_{2:d}^{-1}\Sigma_{2:d,1}$$

We replace the covariances by their rescaled versions to obtain:

$$\tilde{\beta}^{*\top}\widetilde{\Sigma}\tilde{\beta}^* = (D_1\Sigma_{1,2:d}D_{2:d})D_{2:d}^{-1}\Sigma_{2:d}^{-1}D_{2:d}^{-1}(D_{2:d}\Sigma_{2:d,1}D_1) = D_1^2(\beta^*)^\top\Sigma_{2:d}\beta^*$$

For the spiked covariance model, rescaling variances to one amount to rescaling with $D_1 = 1/(1+\theta)$. Thus, we see that our signal strength is affected only by constant factor (assuming $\theta \leq 1$).

4 Experiments

We test our algorithmic framework on randomly generated synthetic data and compare to other existing algorithms for SPCA. The code was implemented in Python using standard libraries.

We refer to our general algorithm from Section 3 that uses the Q statistic as SPCAvSLR. For our SLR "black-box," we use thresholded Lasso [47].[8]. (We experimented with other SLR algorithms such as the forward-backward algorithm of [46] and CoSaMP[9] [32], but results were similar and only slower.)

For more details on our experimental setup, including hyperparameter selection, see Appendix D.

Support recovery We randomly generate a spike $u \in \mathbb{R}^d$ by first choosing a random support of size k, and then using random signs for each coordinate (uniformity is to make sure Condition (C2) is met). Then spike is scaled appropriately with θ to build the spiked covariance matrix of our normal distribution, from which we draw samples.

We study how the performance of six algorithms vary over various values of k for fixed n and d.[10] As in the [15], our measure is the fraction of true support. We compare SPCAvSLR with the following algorithms: diagnoal thresholding, which is a simple baseline; "SPCA" (ZHT [49]) is a fast heuristic also based on the regression idea; the truncated power method of [45], which is known for both strong theoretical guarantees and empirical performance; covariance thresholding, which has state-of-the-art theoretical guarantees.

We modified each algorithm to return the top k most likely coordinates in the support (rather than thresholding based on a cutoff); for algorithms that compute a candidate eigenvector, we choose the top k coordinates largest in absolute value.

We observe that SPCAvSLR performs better than covariance thresholding and diagonal thresholding, but its performance falls short of that of the truncated power method and the heuristic algorithm of [49]. We suspect the playing with different SLR algorithms may slightly improve its performance. The reason for the gap between performance of SPCAvSLR and other state of the arts algorithms despite its theoretical guarantees is open to further investigation.

Hypothesis testing We generate data under two different distributions: for the spiked covariance model, we generate a spike u by sampling a uniformly random direction from the k-dimensional unit sphere, and embedding the vector at a random subset of k coordinates among d coordinates; for the null, we draw from standard isotropic Gaussian. In a single trial, we draw n samples from each distribution and we compute various statistics[11] (diagonal thresholding (DT), Minimal Dual Perturbation (MDP), and our Q statistic, again using thresholded Lasso). We repeat for 100 trials, and plot the resulting empirical distribution for each statistic. We observe similar performance of

---

[8]Thresholded Lasso is a variant of lasso where after running lasso, we keep the k largest (in magnitude) coefficients to make the estimator k-sparse. Proposition 2 in [47] shows that thresholded Lasso satisfies Condition 2.2

[9]CoSaMP in theory requires the stronger condition of restricted isometry on the "sensing" matrix.

[10]We remark that while the size of this dataset might seem too small to be representative "high-dimensional" setting, these are representative of the usual size of dataset that these methods are usually tested on. One bottleneck is the computation of the covariance matrix.

[11]While some of the algorithms used for support recovery in the previous section could in theory be adapted for hypothesis testing, the extensions were immediate so we do not consider them here.
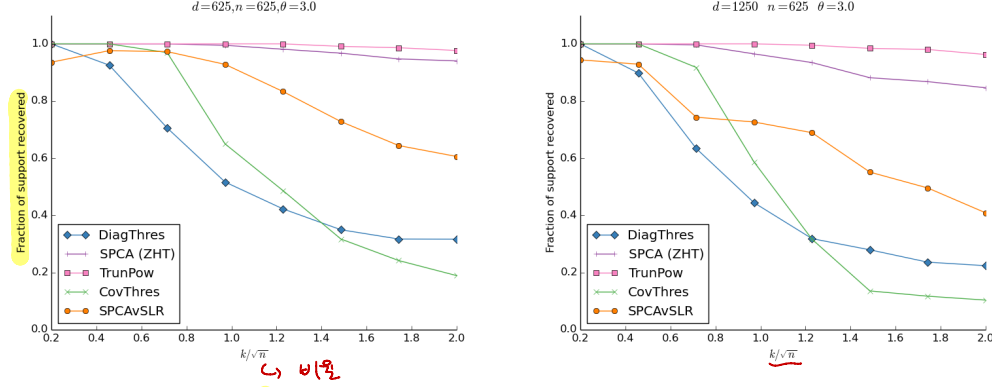
Figure 1: Performance of diagonal thresholding, SPCA (ZHT), truncated power method, covariance thresholding, and SPCAvSLR for support recovery at $n = d = 625$ (left) and $n = 625, d = 1250$ (right), varying values of k, and $\theta = 3.0$. On the horizontal axis we show $k/\sqrt{n}$; the vertical axis is the fraction of support correctly recovered. Each datapoint on the figure is averaged over 50 trials.

DT and Q, while MDP seems slightly more effective at distinguishing H_0 and H_1 at the same signal strength (that is, the distributions of the statistics under H_0 vs. H_1 are more well-separated).

Rescaling variables As discussed in Section 3.4, our algorithms are robust to rescaling the covariance matrix to the correlation matrix. As illustrated in Figure 2 (right), DT fails while Q appears to be still effective for distinguishing hypotheses the same regime of parameters. Other methods such as MDP and CT also appear to be robust to such rescaling (not shown). This suggests that more modern algorithms for SPCA may be more appropriate than diagonal thresholding in practice, particularly on instances where the relative scales of the variables may not be accurate or knowable in advance, but we still want to be able to find correlation between the variables.
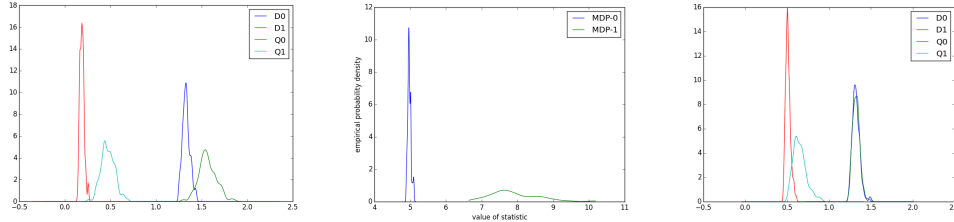

Figure 2: Performance of diagonal thresholding (D), MDP, and Q for hypothesis testing at $n = 200, d = 500, k = 30, \theta = 4$ (left and center). T0 denotes the statistic T under H_0, and similarly for T1. The effect of rescaling the covariance matrix to make variances indistinguishable is demonstrated (right).

5   Previous work

Here we discuss in more detail previous approaches to SPCA and how it relates to our work. Various approaches to SPCA have been designed in an extensive list of prior work. As we cannot cover all of them, we focus on works that aim to give computationally efficient (i.e. polynomial time) algorithms with provable guarantees in settings similar to ours.

These algorithms include fast, heuristic methods based on ℓ_1 minimization [23, 49], rigorous but slow methods based on natural semidefinite program (SDP) relaxations [13, 1, 41, 44], iterative methods motivated by power methods for approximating eigenvectors [45, 24], non-iterative methods based on random projections [20], among others. Many iterative methods rely on initialization schemes, such as ordinary PCA or diagonal thresholding [22].

Below, we discuss the known sample bounds for support recovery and hypothesis testing.

7

Support recovery [1] analyzed both diagonal thresholding and an SDP for support recovery under the spiked covariance model.[12] They showed that the SDP requires an order of k fewer samples when the SDP optimal solution is rank one. However, [27] showed that the rank one condition does not happen in general, particularly in the regime approaching the information theoretic limit ($\sqrt{n} \lesssim k \lesssim n/\log d$). This is consistent with computational lower bounds from [3] ($k \gtrsim \sqrt{n}$), but a small gap remains (diagonal thresholding and SDP's succeed only up to $k \lesssim \sqrt{n/\log d}$). The above gap was closed by the covariance thresholding algorithm, first suggested by [27] and analyzed by [15], that succeeds in the regime $\sqrt{n/\log d} \lesssim k \lesssim \sqrt{n}$, although the theoretical guarantee is limited to the regime when $d/n \to \alpha$ due to relying on techniques from random matrix theory.

Hypothesis testing Some works [4, 1, 17] have focused on the problem of detection. In this case, [4] observed that it suffices to work with the much simpler dual of the standard SDP called Minimal Dual Perturbation (MDP). Diagonal thresholding (DT) and MDP work up to the same signal threshold θ as for support recovery, but MDP seems to outperform DT on simulated data [4]. MDP works at the same signal threshold as the standard SDP relaxation for SPCA. [17] analyze a statistic based on an SDP relaxation and its approximation ratio to the optimal statistic. In the regime where k, n are proportional to d, their statistic succeeds at a signal threshold for θ that is independent of d, unlike the MDP. However, their statistic is quite slow to compute; runtime is at least a high order polynomial in d.

Regression based approaches To the best of our knowledge, our work is the first to give a general framework for SPCA that uses SLR in a *black-box* fashion. [49] uses specific algorithms for SLR such as Lasso as a subroutine, but they use a heuristic alternating minimization procedure to solve a non-convex problem, and hence lack any theoretical guarantees. [31] applies a regression based approach to a restricted class of graphical models. While our regression setup is similar, their statistic is different and their analysis depends directly on the particulars of Lasso. Further, their algorithm requires extraneous conditions on the data.[9] also uses a reduction to linear regression for their problem of sparse subspace estimation. Their iterative algorithm depends crucially on a good initialization done by a diagonal thresholding-like pre-processing step, which fails under rescaling of the data.[13] Furthermore, their framework uses regression for the specific case of orthogonal design, whereas our design matrix can be more general as long as it satisfies a condition similar to the restricted eigenvalue condition. On the other hand, their setup allows for more general ℓ_q-based sparsity as well as the estimation of an entire subspace as opposed to a single component. [29] also achieves this more general setup, while still suffering from the same initialization problem.

Sparse priors Finally, connections between SPCA and SLR have been noted in the probabilistic setting [26, 25], albeit in an indirect manner: the same sparsity-inducing priors can be used for either problem. We view our work as entirely different as we focus on giving a black-box reduction. Furthermore, provable guarantees for the EM algorithm and variational methods are lacking in general, and it is not immediately obvious what signal threshold their algorithm achieves for the single spike covariance model.

6 Conclusion

We gave a black-box reduction for reducing instances of the SPCA problem under the spiked covariance model to instances of SLR. Given oracle access to SLR black-box meeting a certain natural condition, the reduction is shown to efficiently solve hypothesis testing and support recovery.

Several directions are open for future work. The work in this paper remains limited to the Gaussian setting and to the single spiked covariance model. Making the results more general would make the connection made here more appealing. Also, the algorithms designed here, though simple, seem a bit wasteful in that they do not aggregate information from different statistics. Designing a more efficient estimator that makes a more efficient use of samples would be interesting. Finally, there is certainly room for improvement by tuning the choice of the SLR black-box to make the algorithm more efficient for use in practice.

---

[12]They analyze the subcase when the spike is uniform in all k coordinates.

[13]See Section 3.4 for more discussion on rescaling.

References

[1] Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2454–2458. IEEE, 2009.

[2] Afonso S Bandeira, Edgar Dobriban, Dustin G Mixon, and William F Sawin. Certifying the restricted isometry property is hard. *IEEE transactions on information theory*, 59(6):3448–3450, 2013.

[3] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[4] Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[5] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[6] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[7] Florentina Bunea, Alexandre B Tsybakov, and Marten H Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[8] Florentina Bunea, Alexandre B Tsybakov, and Marten H Wegkamp. Sparse density estimation with ℓ_1 penalties. In *Learning theory*, pages 530–543. Springer, 2007.

[9] T Tony Cai, Zongming Ma, and Yihong Wu. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[10] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[11] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[12] Dong Dai, Philippe Rigollet, Lucy Xia, and Tong Zhang. Aggregation of affine estimators. *Electronic Journal of Statistics*, 8(1):302–327, 2014.

[13] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

[14] Gamarnik David and Zadik Ilias. High dimensional regression with binary coefficients. estimating squared error and a phase transtition. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 948–953. PMLR, 2017.

[15] Yash Deshpande and Andrea Montanari. Sparse pca via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–342, 2014.

[16] Thong T Do, Lu Gan, Nam Nguyen, and Trac D Tran. Sparsity adaptive matching pursuit algorithm for practical compressed sensing. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 581–587. IEEE, 2008.

[17] Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui. Approximation bounds for sparse principal component analysis. *Mathematical Programming*, 148(1-2):89–110, 2014.

[18] Alyson K Fletcher, Sundeep Rangan, and Vivek K Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, 2009.

[19] David Gamarnik and Ilias Zadik. Sparse high-dimensional linear regression. algorithmic barriers and a local search algorithm. *arXiv preprint arXiv:1711.04952*, 2017.

[20] Milana Gataric, Tengyao Wang, and Richard J Samworth. Sparse principal component analysis via random projections. *arXiv preprint arXiv:1712.05630*, 2017.

[21] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[22] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 2009.

[23] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.

[24] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb):517–553, 2010.

[25] Rajiv Khanna, Joydeep Ghosh, Russell A Poldrack, and Oluwasanmi Koyejo. Sparse submodular probabilistic pca. In *AISTATS*, 2015.

[26] Oluwasanmi O Koyejo, Rajiv Khanna, Joydeep Ghosh, and Russell Poldrack. On prior distributions and approximate inference for structured variables. In *Advances in Neural Information Processing Systems*, pages 676–684, 2014.

[27] Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations really solve sparse pca. Technical report, Technical report, Weizmann Institute of Science, 2013.

[28] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[29] Zongming Ma et al. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[30] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[31] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[32] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[33] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[34] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[35] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[36] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[37] Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1, 2012.

[38] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[39] Sara van de Geer. The deterministic lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich, 2007.

[40] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[41] Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems*, pages 2670–2678, 2013.

[42] Martin Wainwright. Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, pages 961–965. IEEE, 2007.

[43] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009.

[44] Tengyao Wang, Quentin Berthet, Richard J Samworth, et al. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.

[45] Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *The Journal of Machine Learning Research*, 14(1):899–925, 2013.

[46] Tong Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Advances in Neural Information Processing Systems*, pages 1921–1928, 2009.

[47] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *arXiv preprint arXiv:1402.1918*, 2014.

[48] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *arXiv preprint arXiv:1503.03188*, 2015.

[49] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Appendix

The appendix is organized as follows. In Section A, we give additionaal background on the literature for sparse linear regression. In Section B, we give calculations for quantities from Section 3.1 in the main text, where we set up our main linear model. In Section C, we give complete proofs for Theorems 3.1 and 3.2. Finally, we provide additional details for our experiments in Section D.

A Additional background on SLR

Efficient methods The ℓ_0 estimator, which minimizes the reconstruction error $\|y - \mathbb{X}\widehat{\beta}\|_2^2$ over all k-sparse regression vectors, achieves prediction error bound of form ([7], [35]): $1/n\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2 \lesssim (\sigma^2 k \log d)/n$ but takes exponential time $O(n^k)$ to compute. Various efficient methods have been proposed to circumvent this computational intractability: basis pursuit, Lasso[38], and the Dantzig selector [10] are some of initial approaches. Greedy pursuit methods such as OMP [30], IHT[6], CoSaMP[32], and FoBa[46] among others offer more computationally efficient alternatives.[14] These algorithms achieve the same prediction error guarantee as ℓ_0 up to a constant, but under the assumption that $\mathbb{X}$ satisfies certain properties, such as restricted eigenvalue ([5]), compatibility ([39]), restricted isometry property ([11]), (in)coherence ([8]), among others. In this work, we focus on the restricted eigenvalue (see Definition 2.1 for a formal definition). We remark that restricted eigenvalue is among the weakest, and is only slightly stronger than the compatibility condition. Moreover, [47] give complexity-theoretic evidence for the necessity of dependence on the RE constant for certain worst case instances of the design matrix. See [40] for implications between various conditions. Without such conditions on $\mathbb{X}$, the best known guarantees provably obtain only a $1/\sqrt{n}$ decay rather than a $1/n$ decay in prediction error as number of samples increase; [48] give some evidence that this gap may be unavoidable.

Optimal estimators The SLR estimators we consider are efficiently computable. Another line of work considers arbitrary estimators that are not necessarily efficiently computable. These include BIC [7], Exponential Screening [36], and Q-aggregation [12]. Such estimators achieve strong guarantees regarding minimax optimality in the form of oracle inequalities on MSE.

Restricted Eigenvalue The restricted eigenvalue (RE) lower bounds the quadratic form defined by $\mathbb{X}$ in all (approximately) sparse directions. RE is related to more general notions such as the *restricted strong convexity* [33], which roughly says that loss function is not too "flat" near the point of interest; this allows us to convert convergence in loss value to convergence in parameter value. In general when $d > n$, we cannot guarantee this for all directions, but it suffices to consider the set $\mathbb{C}(S)$ of "mostly" sparse directions.

We remark that the above condition is very natural and likely unavoidable. [47] indicate that the dependence of the above prediction error guarantee on RE cannot be removed, under a standard conjecture in complexity theory. [34, 37] show that RE holds with high probability for correlated Gaussian designs (though it remains NP-hard to verify it [2]).

A recent line of work [14, 19] studies the algorithmic hardness of SLR when $\mathbb{X}$ has Gaussian design.

---

[14]Note that some of these algorithms were presented for compressed sensing; nonetheless, their guarantees can be converted appropriately.

B Deferred calculations from Section 3.1

B.1 Linear minimum mean-square-error estimation

Given random variables Y and X (this can be a vector more generally), what is the best prediction for Y conditioned on knowing $X = x$? What is considered "best" can vary, but here we consider the mean squared error. That is, we want to come up with $\hat{y}(x)$ s.t.

$$\mathbb{E}[(Y - \hat{y})^2]$$

is minimized.

It is not hard to show that $\hat{y}$ is just the conditional expectation of Y conditioned on X. The minimum mean-square-error estimate can be a highly nontrivial function of X.

The *linear minimum mean-square-error* (LMMSE) estimate instead restricts the attention to estimators of the form $\hat{Y} = AX + b$. Notice here that A and b are fixed and are not functions of X.

One can show that the LMMSE estimator is given by: $A = (\Sigma_{XX})^{-1})\Sigma_{XY}$, where Σ. is the appropriately indexed covariance matrix, and b is chosen in the obvious way to make our estimator unbiased.

B.2 Calculations for the linear model

To recap our setup, we input the design matrix $\mathbb{X} = \mathbf{X}_{-i}$ and the response variable $y = \mathbf{X}_i$ as inputs to an SLR black-box. Our goal is to express y as a linear function of $\mathbb{X}$ plus some independent noise w. Without loss of generality let $i = 1$, and for our discussion below assume $S = \{1, ..., k\}$. For illustration, at times we will simplify our calculation further for the uniform case where $u_i = \frac{1}{\sqrt{k}}$ for $1 \le i \le k$ and $u_i = 0$ for $i > k$.

For the moment, just consider one row of $\mathbf{X}$, corresponding to one particular sample X of the original SPCA distribution. Since X is jointly Gaussian, we can express (the expectation of) $y = X_1$ as a linear function of the other coordinates:

$$\mathbb{E}[X_1 | X_{2:d} = x_{2:d}] = \Sigma_{1,2:d}(\Sigma_{2:d})^{-1} x_{2:d}$$

Hence we can write

$$X_1 = \Sigma_{1,2:d}(\Sigma_{2:d})^{-1} X_{2:d} + w$$

where $w \sim \mathcal{N}(0, \sigma^2)$ for some σ to be determined and $w \perp X_i$ for $i = 2, .., d$.

By directly computing the variance of the above expression for X_1, we deduce an expression for the noise level:

$$\sigma^2 = \Sigma_{11} - \Sigma_{1,2:d}(\Sigma_{2:d})^{-1}\Sigma_{2:d,1}$$

Note that σ^2 is just Σ_{11} under H_0. We proceed to compute σ^2 under H_1, when $\Sigma = I_d + \theta u u^\top$. To compute $(\Sigma_{2:d})^{-1}$, we use (a special case of) the Sherman-Morrison formula: $(I + wv^\top)^{-1} = I - \frac{wv^\top}{1+v^\top w}$.

$$\Sigma_{2:d}^{-1} = \left(I_{d-1} + \theta u_{-1} u_{-1}^\top\right)^{-1} = I_{d-1} - \frac{\theta}{1 + (1 - u_1^2)\theta} u_{-1} u_{-1}^\top$$

where $u_{-1} \in \mathbb{R}^{d-1}$ is u restricted to coordinates $2, ..., d$.

$$\begin{aligned}
\Sigma_{1,2:d}(\Sigma_{2:d})^{-1}\Sigma_{2:d,1} &= \left(\frac{\theta u_1}{1 + (1 - u_1^2)}\right)^2 u_{-1}^\top(I + \theta u_{-1} u_{-1}^\top)u_{-1} \\
&= \frac{\theta^2 u_1^2(1 - u_1^2)}{1 + (1 - u_1^2)\theta} \\
&\quad \text{(specializing to uniform case again)} \\
&= \frac{\theta^2}{k}\left(1 - \frac{1}{k}\right)\frac{1}{1 + \frac{k-1}{k}\theta} \approx \frac{\theta^2}{k(1 + \theta)}
\end{aligned}$$

Finally, substituting into the expression for σ^2

$$\sigma^2 = 1 + \theta u_1^2 - \frac{\theta^2 u_1^2 (1 - u_1^2)}{1 + (1 - u_1^2)\theta}$$

$$= 1 + \frac{\theta u_1^2}{1 + (1 - u_1^2)\theta}$$

$$\leq 2 \quad \text{if } \theta \leq 1$$

We remark that the noise level of column 1 has been reduced by roughly $\tau := \frac{\theta^2}{k(1+\theta)}$ by regressing on correlated columns.

In summary, under H_1 (and if $1 \in S$) we can write

$$y = \mathbb{X}\beta^* + w$$

where

$$\beta^* = (\Sigma_{2:d})^{-1}\Sigma_{2:d,1}$$

$$= (I - \frac{\theta}{1 + (1 - u_1)^2\theta}u_{-1}u_{-1}^\top)\theta u_1 u_{-1}$$

$$= \theta u_1 \left(1 - \frac{\theta}{1 + (1 - u_1^2)\theta}(1 - u_1^2)\right)u_{-1}$$

$$= \frac{\theta u_1}{1 + (1 - u_1^2)\theta}u_{-1}$$

(technically, the definition of β^* on the RHS is a $k - 1$ dimensional vector, but we augment it with zeros to make it $d - 1$ dimensional) and $w \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 1 + \frac{\theta u_1^2}{1+(1-u_1^2)\theta}$. Note that in the uniform case, $\beta^* \to \frac{1}{k-1}\mathbf{1}_{k-1}$ as $\theta \to \infty$ where $\mathbf{1}_{k-1}$ is uniform 1 on first $k - 1$ coordinates, as expected.

B.3   Properties of the design matrix $\mathbb{X}$

Restricted eigenvalue (RE)   Here we check that $\mathbb{X}$ defined as in Section 3.1 has constant restricted eigenvalue constant. This allows us to apply Condition 2.2 for the SLR black-box with good guarantee on prediction error.

The rows of $\mathbb{X}$ are drawn from $\mathcal{N}(0, I_{d-1\times d-1} + \theta u_{-1}u_{-1}^\top)$ where u_{-1} is u restricted to coordinates $2, ..., d$ wlog.[15]

Let $\Sigma = I_{d-1\times d-1} + \theta u_{-1}u_{-1}^\top$. We can show that $\Sigma^{1/2}$ satisfies RE with $\gamma = 1$ by bounding Σ's minimum eigenvalue. First, we compute the eigenvalues of $\theta u_{-1}u_{-1}^\top$. $\theta u_{-1}u_{-1}^\top$ has a nullspace of dimension $d - 2$, so eigenvalue 0 has multiplicity $d - 2$. u_{-1} is a trivial eigenvector with eigenvalue $\theta u_{-1}^\top u_{-1} = \theta\frac{k-1}{k}$. Therefore, Σ has eigenvalues 1 and $1 + \theta\frac{k-1}{k}$.

Now we can extend this to the sample matrix $\mathbb{X}$ by applying Corollary 1 of [34] (also see Example 3 therein), and conclude that as soon as $n \gtrsim \frac{\max_j \Sigma_{jj}}{\gamma^2}k \log d = C(1 + \frac{\theta}{k})k \log d = \Omega(k \log p)$ the matrix $\mathbb{X}$ satisfies RE with $\gamma(\mathbb{X}) = 1/8$.

We remark that the following small technical condition also appears in known bounds on prediction error:

Column normalization   This is a condition on the scale of $\mathbb{X}$ relative to the noise in SLR, which is always σ^2.

$$\frac{\|\mathbb{X}\theta\|_2^2}{n} \leq \|\theta\|_2^2$$

---

[15]We assume here that $1 \in S$ as in the previous section

for all $\theta \in B_0(2k)$

We can always rescale $\mathbf{X}$ (and hence $\mathbb{X}$) to satisfy this, which would also rescale the noise level σ in our linear model since the noise is derived from coming $\mathbf{X}$ from the SPCA generative model, rather than added independently as in the usual SLR setup.

Hence, since all scale dependent quantities are scaled by the same amount when we scale the original data $\mathbf{X}$, wlog we may continue to use the same $\mathbb{X}$ and σ in our analysis. As the column normalization condition does not affect us, we drop it from Condition 2.2 of our black-box assumption.

C   Proofs of main Theorems

In this section we analyze the distribution of Q_i under both H_0 and H_1 on our way to proving Theorems 3.1 and 3.2. Note that though the dimension and the sparsity of our SLR instances are $d-1$ and $k-1$ (since we remove one column from the SPCA data matrix $\mathbf{X}$ to obtain the design matrix $\mathbb{X}$), for ease of exposition we just use d, k in their place since it does not affect the analysis in any meaningful way.

First, we review a useful tail bound on χ^2 random variables.

Lemma C.1 (Concentration on upper and lower tails of the χ^2 distribution ([28], Lemma 1)). *Let Z be the χ^2 random variable with k degrees of freedom. Then,*

$$\Pr(Z - k \geq 2\sqrt{kt} + 2t) \leq \exp(-t)$$
$$\Pr(Z - X \geq 2\sqrt{kt}) \leq \exp(-t)$$

We can simplify the upper tail bound as follows for convenience:

Corollary C.2. *For χ^2 r.v. Z with k degrees of freedom and deviation $t \geq 1$, $\Pr\left(\frac{Z-k}{k} \geq 4t\right) \leq \exp(-kt)$.*

C.1   Analysis of Q_i under H_1

Without loss of generality assume the support of u, denoted S, is $\{1, ..., k\}$ and consider the first coordinate. We expand Q_1 by using $y = \mathbb{X}\beta^* + w$ as follows:

$$Q_1 = \frac{1}{n}\|y\|_2^2 - \frac{1}{n}\|y - \mathbb{X}\widehat{\beta}\|_2^2 = \frac{1}{n}\|\mathbb{X}\beta^* + w\|_2^2 - \frac{1}{n}\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2 - \frac{2}{n}w^\top(\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}) - \frac{1}{n}\|w\|_2^2$$

$$= \frac{1}{n}\|\mathbb{X}\beta^*\|_2^2 - \frac{2}{n}w^\top\mathbb{X}\beta^* - \frac{1}{n}(\|\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta}\|_2^2) - \frac{2}{n}w^\top(\mathbb{X}\beta^* - \mathbb{X}\widehat{\beta})$$

Observe that the noise term $\|w\|_2^2$ cancels conveniently.

Before bounding each of these four terms, we introduce a useful lemma to bound cross terms involving noise w:

Lemma C.3 (Lemmas 8 and 9, [35]). *For any fixed $\mathbb{X} \in \mathbb{R}^{n \times d}$ and independent noise vector $w \in \mathbb{R}^n$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries:*

$$\frac{|w^\top\mathbb{X}\theta|}{n} \leq 9\sigma\frac{\|\mathbb{X}\theta\|_2}{n}\sqrt{k\log\frac{d}{k}}$$

for all $\theta \in B_0(2k)$ w.p. at least $\geq 1 - 2\exp(-40k\log(d/k))$

We bound each term as follows:

Term 1. The first term $\frac{\|\mathbb{X}\beta^*\|_2^2}{n}$ contains the signal from the spike; notice its resemblance to the k-sparse eigenvalue statistic. Rewritten in another way,

$$(\beta^*)^\top\frac{\mathbb{X}^\top\mathbb{X}}{n}\beta^* = (\beta^*)^\top\widehat{\Sigma}_{2:d}\beta^*$$

Hence, we expect this to concentrate around $(\beta^*)^\top\Sigma_{2:d}\beta^*$, which simplifies to (see Appendix B.2 for the full calculation):

$$(\beta^*)^\top\Sigma_{2:d}\beta^* = (\Sigma_{1,2:d}\Sigma_{2:d}^{-1})\Sigma_{2:d}(\Sigma_{2:d}^{-1}\Sigma_{2:d,1}) = \frac{\theta^2 u_1^2(1 - u_1^2)}{1 + (1 - u_1^2)\theta}$$

14

For concentration, observe that we may rewrite

$$(\beta^*)^\top \widehat{\Sigma}_{2:d}\beta^* = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}^{(i)}\beta^*)^2$$

where $\mathbb{X}^{(i)}$ is the ith row, representing the ith sample. This is just an appropriately scaled chi-squared random variable with n degrees of freedom (since each $\mathbb{X}^{(i)}\beta^*$ is i.i.d. normal), and the expected value of each term in the sum is the same as computed above. Applying a lower tail bound on χ^2 distribution (see Appendix), with probability at least $1 - \delta$ we have

$$(\beta^*)^\top \widehat{\Sigma}_{2:d}\beta^* \geq \frac{\theta^2 u_1^2(1-u_1^2)}{1+(1-u_1^2)\theta}\cdot\left(1 - 2\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Choosing $\delta = \exp(-k\log d)$,

$$\frac{\|\mathbb{X}\beta^*\|_2^2}{n} \geq \frac{\theta^2 u_1^2(1-u_1^2)}{1+(1-u_1^2)\theta}\cdot\left(1 - 2\sqrt{\frac{k\log d}{n}}\right)$$

$$\overset{(a)}{\geq} \frac{1}{2}\cdot\frac{\theta^2 u_1^2(1-u_1^2)}{1+(1-u_1^2)\theta}$$

$$\overset{(b)}{\geq} \frac{c_{min}^2}{4}\frac{\theta^2}{k} \tag{1}$$

where (a) as long as $n > 16k\log d$ and (b) since $\theta \leq 1$ and $u_1^2(1-u_1^2) \gtrsim c_{min}^2/k$ under Condition (C1).

Term 2. The absolute value of the second term $\frac{2}{n}w^\top \mathbb{X}\beta^*$ can be bounded by $18\frac{\|\mathbb{X}\beta^*\|_2}{n}\sqrt{k\log\frac{d}{k}}$ using Lemma C.3. From (1) as long as $\theta^2 > \frac{c_1}{c_{min}^2}\frac{k^2\log d}{n}$ (c_1 is some constant that we will choose later),

$$\frac{\|\mathbb{X}\beta^*\|_2^2}{n} \geq \frac{c_{min}^2}{4}\frac{\theta^2}{k} \geq \frac{c_1}{4}\frac{k\log d}{n}$$

so the first two terms together are lower bounded by:

$$\frac{\|\mathbb{X}\beta^*\|_2}{n}\left(\|\mathbb{X}\beta^*\|_2 - 18\sqrt{k\log d/k}\right) \geq \frac{c_1}{5}\frac{k\log d}{n}, \tag{2}$$

for large enough constant c_1.

Term 3. The third term, which is the prediction error $\frac{\|\mathbb{X}\beta^* - \mathbb{X}\hat{\beta}\|_2^2}{n}$, is upper bounded by $\frac{C}{\gamma(X)^2}\frac{\sigma^2 k\log d}{n}$ with probability at least $1 - C\exp(-C'k\log d)$ by Condition 2.2 on our SLR black-box. Note $\sigma^2 < 2$ if we assume $\theta \leq 1$.[16] Now, $\gamma(X) \geq \frac{1}{8}$ with probability at least $1 - C\exp(-C'n)$ if $n > C''k\log d$ since $\theta \leq 1$ (see Appendix B.3 for more details). Then,

$$\frac{1}{n}\|\mathbb{X}\beta^* - \mathbb{X}\hat{\beta}\|_2^2 \leq C\frac{k\log d}{n}$$

Term 4. The contribution of the last cross term $\frac{2}{n}w^\top\mathbb{X}(\beta^* - \widehat{\beta})$ can also be bounded by Lemma C.3 w.h.p. (note $\beta^* - \widehat{\beta} \in B_0(2k)$)

$$\frac{|w^\top\mathbb{X}(\beta^* - \widehat{\beta})|}{n} \leq 9\sigma\frac{\|\mathbb{X}(\beta^* - \widehat{\beta})\|_2}{n}\sqrt{k\log\frac{d}{k}}.$$

Combined with the above bound for prediction error, this bounds the cross term's contribution by at most $C'\frac{k\log d}{n}$.

Putting the bounds on four terms together, we get the following lower bound on Q.

---

[16]As smaller θ makes the problem only harder, we assume $\theta \leq 1$ for ease of computation and as standard in literature.

15

Lemma C.4. *There exists constants c_1, c_2, c_3, c_4 s.t. if $\theta^2 > \frac{c_1}{c_{min}^2} \frac{k^2 \log d}{n}$ and $n > c_2 k \log d$, with probability at least $1 - c_3 \exp(-c_4 k \log d)$, for any $i \in S$ that satisfies the size bound in Condition (C1),*

$$Q_i > \frac{13k \log d}{n}$$

Proof. From 1-4 above, by union bound, all four bounds fail to hold with probability at most $c_3 \exp(-c_4 k \log d)$ for appropriate constants if $\theta^2 > \frac{c_1}{c_{min}^2} \frac{k^2 \log d}{n}$ (required by Term 2) and $n > c_2 k \log d$ for some $c_2 > 0$ (note that both terms 1 and 3 require sufficient number of samples n). If all four bounds hold, we have:

$$Q_i > \frac{c_1}{5} \frac{k \log d}{n} - C' \frac{k \log d}{n}$$

where C, C' are just some constants. So if c_1 is sufficiently large, the above bound is greater than $\frac{13k \log d}{n}$.[17] $\square$

C.2   Analysis of Q_i under H_0

We could proceed by decomposing Q_i the same way as in H_1; all the error terms including prediction error are still bounded by $O(k \log d/n)$ in magnitude, and the signal term is gone now since $\beta^* = 0$. This will give the same upper bound (up to a constant) as the following proof is about to show. However, we find the following direct analysis more informative and intuitive.

Since our goal is to upper bound Q_i under H_0, we may let $\widehat{\beta}$ be the optimal possible choice given y and $\mathbb{X}$ (one that minimizes $\|y - \mathbb{X}\widehat{\beta}\|_2^2$, and hence maximizes Q_i). We further break this into two steps. We enumerate over all possible subsets S of size k, and conditioned on each S, choose the optimal $\widehat{\beta}$.

Fix some support S of size k. The span of $\mathbb{X}_S$ is at most a k-dimensional subspace of $\mathbb{R}^n$. Hence, we can consider some unitary transformation U of $\mathbb{R}^n$ that maps the span of $\mathbb{X}_S$ into the subspace spanned by the first k standard basis vectors. Since U is an isometry by definition,

$$nQ_i = \|y\|_2^2 - \|y - \mathbb{X}\widehat{\beta}_S\|_2^2 = \|Uy\|_2^2 - \|Uy - U\mathbb{X}\widehat{\beta}_S\|_2^2$$

Let $\tilde{y} = Uy$. Since $U\mathbb{X}\widehat{\beta}_S$ has nonzero entries only in the first k coordinates, the optimal choice (in the sense of maximizing the above quantity) of $\widehat{\beta}_S$ is to choose linear combinations of the first k columns of $\mathbb{X}$ so that $U\mathbb{X}\widehat{\beta}_S$ equals the first k coordinates of $\tilde{y}$. Then, nQ_i is just the squared norm of the first k coordinates of $\tilde{y}$. Since U is some unitary matrix that is independent of y (being a function of $\mathbb{X}_S$ which is independent of y), $\tilde{y}$ still has i.i.d. $\mathcal{N}(0, 1)$ entries, and hence nQ_i is a χ^2-var with k degrees of freedom.

Now we apply an upper tail bound on the χ^2 distribution (see Corollary C.2). Choosing $t = 3 \log \frac{d}{k}$, and after union bounding over all $\binom{d}{k} \leq \left(\frac{de}{k}\right)^k$ supports S, $nQ_i > k + 12k \log \frac{d}{k} \geq 13k \log \frac{d}{k}$ with probability at most $\exp(-3k \log \frac{d}{k} + k \log \frac{de}{k}) \leq \exp(-k \log \frac{d}{k})$ as long as $\frac{d}{k} \geq e$.

Lemma C.5. *Under H_0, $\forall i$ $Q_i \leq \frac{13k \log \frac{d}{k}}{n}$ w.p. at least $1 - \exp(-k \log \frac{d}{k})$.*

Remark C.6. Union bounding over all S is necessary for the analysis. For instance, we cannot just fix S to be $S(\widehat{\beta})$ (this denotes the support of $\widehat{\beta}$) since $\widehat{\beta}$ is a function of y, so fixing S changes the distribution of y.

Remark C.7. Observe that this analysis of Q_i for H_0 also extends immediately to H_1 when coordinate i is outside the support. The reason the analysis cannot extend to when $i \in S$ is because U is not independent of y in this case.

Corollary C.8. *Under H_1, if $i \notin S$, $Q_i \leq \frac{13k \log \frac{d}{k}}{n}$ w.p. at least $1 - \exp(-k \log \frac{d}{k})$.*

---

[17]The choice of constant 13 may seem a little arbitrary, but this is just to be consistent with Lemma C.5. There, the constant just falls out of convenient choices for simplification, and is not optimized for in particular.

C.3 Proof of Theorem 3.1

Proof. Proof follows immediately from Lemma C.4 and Lemma C.5. We can use our statistics Q_i to separate H_0 and H_1. Under H_0, applying Lemma C.5 to each coordinate i and union bounding, $\forall i$, $Q_i \leq \frac{13k \log \frac{d}{k}}{n}$ with probability at least $1 - \exp(-Ck \log d)$. Meanwhile, under H_1, if we consider any coordinate i that satisfies Condition (C1), Lemma C.4 gives

$$Q_i > \frac{13k \log d}{n}$$

with probability at least $1 - c_3 \exp(-c_4 k \log d)$. Since ψ tests whether $Q_i > \frac{13k \log \frac{d}{k}}{n}$ for at least one i, ψ distinguishes H_0 and H_1 successfully, with bound on type I and type II error probability $c_3 \exp(-c_4 k \log d)$ for appropriate constants c_3, c_4 (note, these may be different from those of Lemma C.4). For runtime, note that we make d calls to the SLR black-box and work with matrices of size $n \times d$. $\square$

C.4 Proof of Theorem 3.2

Proof. As long as every u_i for $i \in S$ has magnitude $c_{min}/\sqrt{k}$ as in Condition (C2), we can repeat the same analysis from above to all coordinates in the support. If θ meets the same threshold, $Q_i > 13k \log \frac{d}{k}/n$ for all $i \in S$ with probability at least $1 - C \exp(-C'k \log d)$ by union bound. Also, recall $Q_i > 13k \log \frac{d}{k}/n$ for any $i \notin S$ with probability at most $C \exp(-C'k \log d)$ by Corollary C.8. By union bound over all $d - k$ coordinates outside the support, the error probability is at most $d \cdot C \exp(-C'k \log d) \leq C \exp(-C''k \log d)$. We showed that with high probability we exactly recover the support S of u.

Runtime analysis is identical to that for the hypothesis test. $\square$

Running time The runtime of both Algorithms 1 and 2 is $\tilde{O}(nd^2)$,[18] if we assume the SLR black-box takes nearly linear time in input size, $\tilde{O}(nd)$, which is achieved by known existing algorithms. Note that computing the sample covariance matrix alone takes $O(nd^2)$ time, assuming one is using a naive implementation of matrix multiplication. For a broad comparison, we consider spectral methods and SDP-based methods, though there are methods that do not fall in either category. Spectral methods such as covariance thresholding or truncated power method have an iteration cost $O(d^2)$ due to operating on $d \times d$ matrices, and hence have total running time $\tilde{O}(d^2)$ ($\tilde{O}(\cdot)$ hiding precise convergence rate) in addition to the same $O(nd^2)$ initialization time. SDP-based methods in general take $\tilde{O}(d^3)$ time, the time taken by interior point methods to optimize. So overall, Algorithms 1 and 2 are competitive choices for (single spiked) SPCA, at least theoretically, though they seem slower in practice.

D Experimental Setup

We provide some futher details of the experimental setup, including selection of hyperparameters.

For the "SPCA" algorithm of [49], we used their direct implementation using an initialization with PCA (rather than the self-contained alternating minimization algorithm they present as an alternative). We also leave out the ℓ_2 ridge penalty for convenience of implementation (their algorithm already performed very well in our experiments, so it was unnecessary to implement the full version).

For the truncated power method, we used the convergence criterion that the l_2 norm of the difference between eigenvectors from two consecutive iterations is less than $\epsilon = 0.01$.

For covariance thresholding, we tried various levels of parameter τ, which controls the threshold for soft-thresholding, and indeed it performed best at [15]'s recommended value of $\tau \approx 4$, which is the choice compared against.

For our Q-based algorithm SPCAvSLR, we used thresholded Lasso with $\lambda = 0.1$ where λ controls the weight of the ℓ_1 regularization. This is close to the recommended choice of $\lambda = 4\sigma \sqrt{\frac{\log d}{n}}$ from [47] for our parameter setting.

---

[18]In what follows $\tilde{O}(\cdot)$ hides possible log and accuracy parameter ε factors.