Sparse PCA Beyond Covariance Thresholding

Gleb Novikov GLEB.NOVIKOV@INF.ETHZ.CH
ETH Zurich

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

In the Wishart model for sparse PCA we are given n samples $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ drawn independently from a d-dimensional Gaussian distribution $N(0, \mathrm{Id} + \beta vv^\top)$, where $\beta > 0$ and $v \in \mathbb{R}^d$ is a k-sparse unit vector, and we wish to recover v (up to sign).

We show that if $n \geqslant \Omega(d)$, then for every $t \ll k$ there exists an algorithm running in time $n \cdot d^{O(t)}$ that solves this problem as long as

$$\beta \gtrsim \frac{k}{\sqrt{nt}} \sqrt{\ln(2 + td/k^2)}\,.$$

Prior to this work, the best polynomial time algorithm in the regime $k \approx \sqrt{d}$, called *Covariance Thresholding* (proposed in Krauthgamer et al. (2015) and analyzed in Deshpande and Montanari (2014)), required $\beta \gtrsim \frac{k}{\sqrt{n}} \sqrt{\ln(2 + d/k^2)}$. For large enough constant t our algorithm runs in polynomial time and has better guarantees than Covariance Thresholding. Previously known algorithms with such guarantees required quasi-polynomial time $d^{O(\log d)}$.

Our idea is based on the idea of Alon et al. (1998) for reducing the clique size in the planted clique problem. Moreover, we show that it is possible to combine our techniques with recent results on sparse PCA with symmetric heavy-tailed noise d'Orsi et al. (2022). Their model generalizes both sparse PCA and the planted clique problem. In particular, in the regime $k \approx \sqrt{d}$ we get the first polynomial time algorithm that works with symmetric heavy-tailed noise, while the algorithm from d'Orsi et al. (2022) requires quasi-polynomial time in these settings. As a consequence, we get an algorithm that solves a problem that captures both sparse PCA and planted clique and achieves best known guarantees for both of them.

In addition, we show that our techniques work with sparse PCA with adversarial perturbations studied in d'Orsi et al. (2020). This model generalizes not only sparse PCA, but also the sparse planted vector problem. As a consequence, we provide polynomial time algorithms for the sparse planted vector problem that have better guarantees than the state of the art in some regimes.

Keywords: Sparse PCA, Adversarial Perturbations, Semidefinite Programming, Symmetric Noise

1. Introduction

We study sparse principal component analysis in the *Wishart* and *Wigner* models. First we describe the Wishart model (that is sometimes also called the *spiked covariance model*). In this model, we are given n samples $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ drawn[1] independently from d-dimensional Gaussian distribution $N\big(0, \mathrm{Id} + \beta vv^\top\big)$, where $\beta > 0$ and $v \in \mathbb{R}^d$ is a k-sparse[2] unit vector. The goal is to compute an estimator $\hat{\boldsymbol{v}}$ such that $\|\hat{\boldsymbol{v}}\| = 1$ and $|\langle \hat{\boldsymbol{v}}, v\rangle|$ is close to 1 (say, is greater than 0.99) with high probability[3]. In this paper we mostly focus on the regime when the number of samples n is greater than the dimension d, and in this section of the paper we always assume that[4] $n \geqslant \Omega(d)$ (unless stated otherwise).

1. We use boldface to denote random variables.
2. That is, this vector has at most k non-zero coordinates.
3. It is impossible to recover the sign of v from $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$.
4. We hide absolute constant multiplicative factors using the standard notations $O(\cdot), \Omega(\cdot), \lesssim, \gtrsim$.

Classical settings. The standard approach in covariance estimation is to consider the empirical covariance $\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}$ (where $\boldsymbol{Y}$ is the matrix with rows $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_n$). The top eigenvector of $\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}$ is highly correlated with v or $-v$ as long as $\beta \gtrsim \sqrt{\frac{d}{n}}$, and in non-sparse settings ($k = d$) these guarantees are information theoretically optimal. For $k < d$, there exists an estimator with better guarantees. It uses exhaustive search over all $\binom{d}{k}$ candidates for the support of v and is close to v or $-v$ iff $\beta \gtrsim \sqrt{\frac{k\log(de/k)}{n}}$, and these guarantees are information theoretically optimal in sparse settings (Amini and Wainwright, 2009; Berthet and Rigollet, 2013b,c).

As was observed in Johnstone and Lu (2009), known algorithmic guarantees for sparse PCA are strictly worse than the statistical guarantees described above. In the regime $k \gg \sqrt{d}$, no polynomial time algorithm is known to work if $\beta \lesssim \sqrt{\frac{d}{n}}$ (recall that if $\beta \gtrsim \sqrt{\frac{d}{n}}$, the top eigenvector of the empirical covariance is a good estimator). Johnstone and Lu (2009) proposed a polynomial time algorithm (called *Diagonal Thresholding*) that finds an estimator that is close to v or $-v$ as long as $\beta \gtrsim k\sqrt{\frac{\log d}{n}}$, which is better than the top eigenvector of $\boldsymbol{Y}^\top\boldsymbol{Y}$ if $k \ll \sqrt{d}$, but is worse than the information-theoretically optimal estimator by a factor $\sqrt{k}$.

Later many computational lower bounds of different kind appeared: reductions from the planted clique problem (Berthet and Rigollet, 2013a,b; Wang et al., 2016; Gao et al., 2017; Brennan et al., 2018; Brennan and Bresler, 2019), low degree polynomial lower bounds Ding et al. (2019); d'Orsi et al. (2020), statistical query lower bounds (Brennan et al., 2021), SDP and sum-of-squares lower bounds (Krauthgamer et al., 2015; Ma and Wigderson, 2015; Potechin and Rajendran, 2022), lower bounds for Markov chain Monte Carlo methods (Arous et al., 2020). These lower bounds suggest that the algorithms described above should have optimal guarantees in the regimes $k \ll \sqrt{d}$ (Diagonal Thresholding) and $k \gg \sqrt{d}$ (the top eigenvector), so it is unlikely that there exist efficient algorithms with significantly better guarantees if $k \ll \sqrt{d}$ or $k \gg \sqrt{d}$.

The regime $k \approx \sqrt{d}$ is more interesting. For a long time no efficiently computable estimator with provable guarantees better than the top eigenvector of $\boldsymbol{Y}^\top\boldsymbol{Y}$ or than Diagonal Thresholding was known, until Deshpande and Montanari (2014) proved that a polynomial time algorithm (called *Covariance Thresholding*) computes an estimator that is close to v or $-v$ as long as $\beta \gtrsim k\sqrt{\frac{\log(2+d/k^2)}{n}}$. This estimator can exploit sparsity if $k < \sqrt{d}$ and is better than Diagonal Thresholding and the top eigenvector of the empirical covariance in the regime $d^{1/2-o(1)} < k < \sqrt{d}$.

These results show that in order to work with smaller signal strength β, one needs either to work with larger number of samples n, or to work with a sparser vector v (i.e. smaller k). Ding et al. (2019) (and independently Holtzman et al. (2020)) showed that in some regimes there is another option: one can (smoothly) increase the running time needed to compute the estimator in order to work with smaller signal strength. Concretely, they showed that for $1 \leqslant t \leqslant k/\log d$ there exists an estimator that can be computed in time $d^{O(t)}$ (via *limited brute force*) and is close to v or $-v$ as long as $\beta \gtrsim k\sqrt{\frac{\log d}{tn}}$. The following example illustrates their result: For some $n, d, k \in \mathbb{N}$, let β_{DT} be the smallest signal strength such that Diagonal Thresholding, given an instance $\boldsymbol{Y}$ of sparse PCA with n samples, dimension d, sparsity k and signal strength β_{DT}, finds a unit vector $\hat{\boldsymbol{v}}_{\text{DT}}$ such that $|\langle\hat{\boldsymbol{v}}_{\text{DT}}, v\rangle| \geqslant 0.99$ with high probability. Now suppose that for the same n, d, k, we are given an instance $\boldsymbol{Y}'$ of sparse PCA with smaller signal strength $\beta_{\text{new}} = 0.01 \cdot \beta_{\text{DT}}$. Then their result implies that there exists a polynomial time algorithm that, given $\boldsymbol{Y}'$, finds a unit vector $\hat{\boldsymbol{v}}_{\text{new}}$ such that $|\langle\hat{\boldsymbol{v}}_{\text{new}}, v\rangle| \geqslant 0.99$ with high probability.

However, the approach of Ding et al. (2019) and Holtzman et al. (2020) is not compatible with the optimal guarantees in the regime $k \approx \sqrt{d}$. More precisely, if we define β_{CT} as the smallest signal strength for Covariance Thresholding (in the same way as we defined β_{DT} for Diagonal Thresholding), then the limited brute force that works with signal strength $0.01 \cdot \beta_{\text{CT}}$ requires t to be at least $\log d$ and hence runs in quasi-polynomial time $d^{O(\log d)}$. Prior to this work it was the fastest algorithm in this regime.

Our result shows that it is possible to smoothly increasy running time in order to work with smaller signal strength as long as $k \leqslant O(\sqrt{d})$ and $n \geqslant \Omega(d)$. It can be informally described as follows: Let $\mathcal{A}$ be an arbitrary currently known polynomial time algorithm for sparse PCA. Let $\beta_{\mathcal{A}}$ be the smallest signal strength such that $\mathcal{A}$, given an instance $\boldsymbol{Y}$ of sparse PCA with, dimension d, $n \geqslant \Omega(d)$ samples, sparsity $k \leqslant O\Big(\sqrt{d}\Big)$, and signal strength $\beta_{\mathcal{A}}$, finds a unit vector $\hat{\boldsymbol{v}}_{\mathcal{A}}$ such that $|\langle \hat{\boldsymbol{v}}_{\mathcal{A}}, v\rangle| \geqslant 0.99$ with high probability. For arbitrary constant $C \geqslant 1$, let $\beta_C = \frac{1}{C}\beta_{\mathcal{A}}$. Then there exists a polynomial time[5] algorithm, that, given an instance of sparse PCA $\boldsymbol{Y}'$ with signal strength β_C and the same parameters n, d, k as for $\boldsymbol{Y}$, finds a unit vector $\hat{\boldsymbol{v}}_{\text{new}}$ such that $|\langle \hat{\boldsymbol{v}}_{\text{new}}, v\rangle| \geqslant 0.99$ with high probability.

In particular, our result implies that there exists a polynomial time algorithm that works with signal strength $0.01 \cdot \beta_{\text{CT}}$, which is a significant improvement compared to the best previously known (quasi-polynomial time) algorithm. Moreover, our result also implies the first polynomial time algorithm that can exploit sparsity and has better guarantees than the top eigenvector even in the regime $k \geqslant \sqrt{d}$ (as long as $k \leqslant O\Big(\sqrt{d}\Big)$).

Semidefinite programming and adversarial perturbations. d'Aspremont et al. (2004) introduced *basic SDP* for sparse PCA. basic SDP achieves guarantees of both the top eigenvector of the empirical covariance and Diagonal Thresholding. Later d'Orsi et al. (2020) proved that it also achieves the guarantees of Covariance Thresholding, and hence captures the best currently known polynomial time guarantees.

Moreover, d'Orsi et al. (2020) showed that basic SDP also works with adversarial perturbations. More precisely, if a small (adversarially chosen) value E_{ij} is added to every entry $\boldsymbol{Y}_{ij}$ of an instance of sparse PCA, basic SDP still recovers v or $-v$ with high probability. Known estimators that are not based on semidefinite programming, including top eigenvector of the empirical covariance, Diagonal Thresholding, Covariance Thresholding and limited brute force, do not work with adversarial perturbations. d'Orsi et al. (2020) also provided a family of algorithms based on sum-of-squares relaxations that work with adversarial perturbations and achieves the guarantees of limited brute force from Ding et al. (2019).

Similar to non-adversarial case, basic SDP and limited brute force based on sum-of-squares are not compatible with each other: in the regime $k \approx \sqrt{d}$, the sum-of-squares approach from d'Orsi et al. (2020) requires degree $\log d$ in order to achieve better guarantees than basic SDP, so the corresponding estimator can be computed only in quasi-polynomial time.

We show that our technique also works with adversarial perturbations. We remark that we do not use higher degree sum-of-squares, but only basic SDP for sparse PCA (with some preprocessing and postprocessing steps).

One of the applications of our result is an improvement in the planted sparse vector problem. For this problem we focus on the regime $\Omega(d) < n < d$. In this problem, we are given an n-dimensional

5. The degree of the polynomial depends on C.

subspace of $\mathbb{R}^d$ that contains a sparse vector, and the goal is to estimate this vector. This problem was extensively studied in literature in different settings (Hand and Demanet, 2013; Barak et al., 2014; Hopkins et al., 2016; Qu et al., 2020; Mao and Wein, 2021; Zadik et al., 2022; Diakonikolas and Kane, 2022). It is not hard to see[6] that this problem in the *Gaussian basis* model (in the sense of Mao and Wein (2021)) is a special case of sparse PCA with small perturbations. Our result shows that as long as $k \leqslant \sqrt{td}$, there exists a $d^{O(t)}$ time algorithm for this problem. Previously known polynomial time algorithms in the regime $n \geqslant \Omega(d)$ required $k \leqslant C\sqrt{d}$ for some absolute constant C and did not work for $k > C\sqrt{d}$. Lower bounds against restricted computational models (d'Orsi et al., 2020; Ding et al., 2021; Ding and Hua, 2023) suggest that in the regime $n \geqslant \Omega(d)$ this problem is unlikely to be solvable in polynomial time if $k \gg \sqrt{d}$.

The Wigner model and symmetric noise. Our results can be naturally applied also to the Wigner model. In this model, we are given $\boldsymbol{Y} = \lambda vv^\top + \boldsymbol{W}$, where $\lambda > 0$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, and $\boldsymbol{W} \sim N(0,1)^{d\times d}$. For this model, Covariance Thresholding finds an estimator highly correlated with with v or $-v$ as long as $\lambda \gtrsim k\sqrt{\log(2+d/k^2)}$, while limited brute force from Ding et al. (2019) computes in time $d^{O(t)}$ an estimator close to v or $-v$ as long as $\lambda \gtrsim k\sqrt{\frac{\log d}{t}}$. As in the Wishart model, these approaches are not compatible in the regime $k \approx \sqrt{d}$. Our techniques can be naturally applied to the Wigner model, leading to the best known algorithms for this problem.

As in the Wishart model, our techniques also work with adversarial perturbations. Moreover, our approach is compatible with the recent study of Sparse PCA with symmetric noise d'Orsi et al. (2022). In this model, Gaussian noise $\boldsymbol{W}$ is replaced by an arbitrary noise $\boldsymbol{N}$ with symmetric about zero independent entries that are only guaranteed to be bounded by 1 with probability[7] $\Omega(1)$. They proposed a quasi-polynomial algorithm for sparse PCA in these settings and provided evidence that in the regime $k \ll \sqrt{d}$ this running time cannot be improved (via reduction from the planted clique problem). Combining their algorithm with our approach, we show that in the regime $k \approx \sqrt{d}$ there exists a polynomial time algorithm that solves this problem.

1.1. Results

Before stating the results, observe that one can write an instance $\boldsymbol{Y}$ sparse PCA in the Wishart model as $\boldsymbol{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W}$, where $u \sim N(0,1)^n$ and $\boldsymbol{W} \sim N(0,1)^{n\times d}$ are independent.

Classical settings. Our first result is estimating v in the Wishart model in classical settings (without perturbations).

Theorem 1 (The Wishart model) *Let $n, d, k, t \in \mathbb{N}$, $\beta > 0$. Let $\boldsymbol{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W}$, where $\boldsymbol{u} \sim N(0,1)^n$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n\times d}$ independent of $\boldsymbol{u}$.*

There exists an absolute constant $C > 1$, such that if $n \geqslant Ck$, $k \geqslant Ct\log^2 d$ and

$$\beta \geqslant C\frac{k}{\sqrt{tn}}\sqrt{\log\left(2+\frac{td}{k^2}\left(1+\frac{d}{n}\right)\right)},$$

then there exists an algorithm that, given $\boldsymbol{Y}$, k and t, in time $n\cdot d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1-o(1)$ as $d\to\infty$,

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 0.99\,.$$

6. See the discussion before corollary 3.
7. Note that even the first moment is not required to exist.

Let us compare our guarantees with previously known estimators. For simplicity we assume $n \geqslant \Omega(d)$. For this regime, best estimators known prior to this work and their guarantees are listed in Table 1.

Estimator	Signal Strength	Time Complexity
Statistically optimal estimator	$\beta \gtrsim \sqrt{\frac{k}{n}\log(ed/k)}$	$n \cdot d^{O(k)}$
Top eigenvector of the empirical covariance	$\beta \gtrsim \sqrt{\frac{d}{n}}$	$n \cdot d^{O(1)}$
Covariance Thresholding	$\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log(2+d/k^2)}$	$n \cdot d^{O(1)}$
Limited brute force from Ding et al. (2019)	$\beta \gtrsim \frac{k}{\sqrt{tn}}\sqrt{\log d}$	$n \cdot d^{O(t)}$
Our estimator	$\beta \gtrsim \frac{k}{\sqrt{tn}}\sqrt{\log(2+td/k^2)}$	$n \cdot d^{O(t)}$

Table 1: Estimators for sparse PCA in the Wishart model (assuming $n \geqslant \Omega(d)$ and $t \leqslant k/\operatorname{polylog}(d)$).

For $k \leqslant d^{1/2-\Omega(1)}$ (say, $k \leqslant d^{0.49}$), the guarantees of the algorithm from Ding et al. (2019) are similar to ours (up to a constant factor). For $k \geqslant d^{1/2-o(1)}$ our algorithm can work with asymptotically smaller signal strength (with the same running time).

To compare with Covariance Thresholding and the top eigenvector of the empirical covariance, consider the regime $k = \Theta\left(\sqrt{d}\right)$. Note in this regime both Covariance Thresholding and the top eigenvector require

$$\beta \geqslant c\sqrt{d/n}$$

for some specific constant c (that depends on $\sqrt{d}/k$), and they do not work for smaller β. Our condition on β in these settings is

$$\beta \gtrsim \frac{k}{\sqrt{tn}}\sqrt{\log t}\,,$$

so if $\beta = \varepsilon\sqrt{d/n}$ for arbitrary constant ε, we can choose large enough constant t such that $\varepsilon\sqrt{d} \gtrsim k\sqrt{\frac{\log t}{t}}$ and get an estimator that is highly correlated with v or $-v$ in polynomial time $n \cdot d^{O(t)}$. Neither Covariance Thresholding nor the top eigenvector of the empirical covariance can work with small values of ε, and limited brute force from Ding et al. (2019) requires quasi-polynomial time $n \cdot d^{O(\log d)}$ in these settings.

It is also interesting to compare our upper bound with the low degree polynomial lower bound from d'Orsi et al. (2020). They showed that in the regime $k \leqslant O\left(\sqrt{d}\right)$, polynomials of degree $D \leqslant n/\log^2 n$ cannot distinguish[8] $\boldsymbol{Y}_1 \ldots, \boldsymbol{Y}_n \sim N(0, \mathrm{Id} + \beta vv^\top)$ from $\boldsymbol{Y}_1 \ldots, \boldsymbol{Y}_n \sim N(0, \mathrm{Id})$ if

$$\beta \lesssim \frac{k}{\sqrt{Dn}} \cdot \log\left(2 + \frac{Dd}{k^2}\right).$$

8. More precisely, they cannot *strongly distinguish* sequences of distributions in the sense of Kunisky et al. (2019).

and hence for such β they cannot be used to design an estimator that is close to v or $-v$ with high probability. Their lower bound does not formally imply that our upper bound is tight (that is, it does not imply that there are no better estimators than ours among low degree polynomials). However, there is an interesting similarity between the lower bound and the upper bound: They have a very similar logarithmic factor. If this similarity can be formalized, it may lead to an algorithm that works in a small sample regime $n \ll d$. The term d/n that we have in the logarithmic factor in the bound on β is necessary for our techniques. Many other algorithms, like basic SDP or Covariance Thresholding, also have similar terms. However, the low-degree lower bound does not have this term and d'Orsi et al. (2020) provided an algorithm based on low degree polynomials that does not have such a term and works as long as $\beta \gtrsim \frac{k}{\sqrt{n}}\sqrt{\log(2+td/k^2)}$ even for very small n (e.g. $n = d^{0.01}$). Finding an estimator with guarantees similar to ours in the small sample regime $n \ll d$ is an interesting open question, and low degree polynomials might be useful in designing such an estimator.

Adversarial perturbations. Our approach also works in the presence of adversarial perturbations.

Theorem 2 (The Wishart model with adversarial perturbations) *Let $n, d, k, t \in \mathbb{N}$, $\beta > 0$, $\varepsilon \in (0,1)$. Let $Y = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W} + E$, where $\boldsymbol{u} \sim N(0,1)^n$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n\times d}$ independent of $\boldsymbol{u}$ and $E \in \mathbb{R}^{n\times d}$ is a matrix such that*

$$\|E\|_{1\to 2} \leqslant \varepsilon \cdot \min\Big\{\sqrt{\beta}, \beta\Big\} \cdot \sqrt{n/k}\,,$$

where $\|E\|_{1\to 2}$ is the maximal norm of the columns of E and $\varepsilon < 1$.

There exists an absolute constant $C > 1$, such that if $n \geqslant Ck$, $k \geqslant Ct\log^2 d$,

$$\beta \geqslant C\frac{k}{\sqrt{tn}}\sqrt{\log\left(2 + \frac{td}{k^2}\left(1+\frac{d}{n}\right)\right)}\,.$$

and $\varepsilon\sqrt{\log(1/\varepsilon)} \leqslant \frac{1}{C}\min\Big\{1, \min\big\{\beta, \sqrt{\beta}\big\}\cdot\sqrt{n/d}\Big\}$, then there exists an algorithm that, given Y, k and t, in time $n\cdot d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1-o(1)$ as $d\to\infty$,

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 0.99\,.$$

To illustrate how large the adversarial perturbations are allowed to be, consider the following example: Let $k = \Theta(\sqrt{d})$, $n = \Theta(d)$, $\beta = \Theta(1)$, $t \leqslant O(1)$. Then the columns of E can have norm as large as $\Omega\Big(\sqrt{k}\Big)$. Note that in these settings column norms of $\sqrt{\beta}uv^\top$ can be $O\Big(\sqrt{k}\Big)$. Hence, in this regime, if we allow E to be larger by a constant factor, the adversary can choose $E = -\sqrt{\beta}uv^\top$ and erase the signal. As was shown in d'Orsi et al. (2020), in these settings Covariance Thresholding, Diagonal Thresholding and the top eigenvector of the empirical covariance do not work with some perturbations E such that $\|E\|_{1\to 2} \leqslant k^{o(1)}$.

Our assumption on E is stronger than the assumption from d'Orsi et al. (2020), which is $\|E\|_{1\to 2} \lesssim \min\big\{\beta, \sqrt{\beta}\big\}\sqrt{n/k}$. Designing an estimator with guarantees similar to ours that works with larger E is an interesting problem.

Similar to the non-adversarial settings, our algorithms have the same guarantees[9] as the sum-of-squares approach from d'Orsi et al. (2020) if $k \leqslant d^{1/2-\Omega(1)}$ and has asymptotically better guarantees

9. Assuming our bound on the columns of E.

if $k \geqslant d^{1/2-o(1)}$. Similarly to Covariance Thresholding in the non-adversarial case, in the regime $n \geqslant \Omega(d)$ and $k = \Theta\left(\sqrt{d}\right)$ basic SDP requires $\beta \geqslant c\sqrt{d/n}$ for some specific constant c and does not work for smaller β. Our condition on β in these settings is $\beta \gtrsim \frac{k}{\sqrt{tn}}\sqrt{\log t}$, so if $\beta = \varepsilon\sqrt{d/n}$ for arbitrary constant ε, we can choose large enough constant t such that $\varepsilon\sqrt{d} \gtrsim k\sqrt{\frac{\log t}{t}}$ and get an estimator that can be computed in polynomial time $n \cdot d^{O(t)}$. basic SDP cannot work with small values of ε, and sum-of-squares approach from d'Orsi et al. (2020) requires quasi-polynomial time in these settings.

The sparse planted vector problem. As was observed in d'Orsi et al. (2020), sparse PCA with perturbations is a generalization not only for the spiked covariance model, but also for the planted sparse vector problem. In this problem we are given an n-dimensional subspace of $\mathbb{R}^d$ spanned by $n-1$ random vectors and a sparse vector, and the goal is to find the sparse vector. More precisely, let $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n$ be standard d-dimensional Gaussian vectors and let $\boldsymbol{B}$ be an $n \times d$ matrix whose first $n-1$ rows are $\boldsymbol{g}_1^\top, \ldots, \boldsymbol{g}_{n-1}^\top$ and the last row is a vector $\|\boldsymbol{g}_n\| v^\top$, where $v \in \mathbb{R}^d$ is k-sparse and unit. Let $\boldsymbol{R}$ be a random rotation of $\mathbb{R}^n$ independent of $\boldsymbol{g}_1 \ldots, \boldsymbol{g}_n$, and let $\boldsymbol{Y} = \boldsymbol{RB}$. The goal is to recover v from $\boldsymbol{Y}$.

This problem can be seen as a special case of sparse PCA with perturbation matrix $E = -\frac{1}{\|\boldsymbol{u}\|^2}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{W}$ (see section B for the proof). Therefore, we can apply theorem 2 and get

Corollary 3 (The sparse planted vector problem) *Let $n, d, k, t \in \mathbb{N}$, $\beta > 0$. Let $\boldsymbol{Y} = \sqrt{\boldsymbol{\beta}}\boldsymbol{u}v^\top + \boldsymbol{W} - \frac{1}{\|\boldsymbol{u}\|^2}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{W}$, where $\boldsymbol{u} \sim N(0,1)^n$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n\times d}$ independent of $\boldsymbol{u}$, and $\sqrt{\boldsymbol{\beta}} = \frac{\|\boldsymbol{u}^\top\boldsymbol{W}\|}{\|\boldsymbol{u}\|^2}$.*

There exists an absolute constant $C > 1$, such that if $d > n$, $n \geqslant Ck$, $k \geqslant Ct\log^2 d$ and

$$k \leqslant \frac{1}{C} \cdot d\sqrt{t/n}\,,$$

then there exists an algorithm that, given $\boldsymbol{Y}$, k and t, in time $d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 0.99\,.$$

Prior to this work, in the regime $n \geqslant \Omega(d)$, polynomial time estimators were known only if $k \leqslant cd/\sqrt{n}$ for some small constant $c < 1$ (the existence of such an algorithm follows from Theorem 4.5 from d'Orsi et al. (2020)). We show that even if $k \geqslant 100d/\sqrt{n}$, sparsity can still be exploited and there are estimators that can be computed in polynomial time.

The Wigner model and symmetric noise. Our techniques also work with sparse PCA in the Wigner model.

Theorem 4 (The Wigner model) *Let $k, d, t \in \mathbb{N}$, $\lambda > 0$. Let $Y = \lambda vv^\top + \boldsymbol{W} + E$, where $v \in \mathbb{R}^d$ is a k-sparse unit vector and $\boldsymbol{W} \sim N(0,1)^{d\times d}$ and $E \in \mathbb{R}^{d\times d}$.*

There exists an absolute constant $C > 1$, such that if $k \geqslant Ct\log d$, $\|E\|_\infty \leqslant \frac{1}{C}\lambda/k$, and

$$\lambda \geqslant Ck\sqrt{\frac{\log(2 + td/k^2)}{t}}\,,$$

then there exists an algorithm that, given Y, k and t, in time $d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v \rangle| \geqslant 0.99 \,.$$

Note that E is allowed to be as large as possible (up to a constant factor). Similar to the Wishart model, previously known polynomial time algorithms required

$$\lambda \gtrsim \min\left\{ k\sqrt{\frac{\log d}{t}}, k\sqrt{\log(2 + d/k^2)}, \sqrt{d} \right\},$$

where $k\sqrt{\frac{\log d}{t}}$ corresponds to the limited brute force from Ding et al. (2019), and $\min\left\{k\sqrt{\log(2 + d/k^2)}, \sqrt{d}\right\}$ corresponds to basic SDP. Similarly to the Wishart model, in the regime $k \geqslant d^{1/2 - o(1)}$ we get asymptotically better guarantees than the algorithm from Ding et al. (2019). In the regime $n \geqslant \Omega(d)$ and $k = \Theta\left(\sqrt{d}\right)$ basic SDP requires $\lambda \geqslant c\sqrt{d}$ for some specific constant c and does not work for smaller λ. Our condition on λ in these settings is $\lambda \gtrsim k\sqrt{\frac{\log t}{t}}$, so if $\lambda = \varepsilon\sqrt{d}$ for arbitrary constant ε, we can choose large enough constant t such that $\varepsilon\sqrt{d} \gtrsim k\sqrt{\frac{\log t}{t}}$ and get an estimator that can be computed in polynomial time $d^{O(t)}$. basic SDP cannot work with small values of ε, and limited brute force requires quasi-polynomial time in these settings.

Our techniques can be also applied to more general model with symmetric noise studied in d'Orsi et al. (2022).

Theorem 5 (Sparse PCA with symmetric heavy-tailed noise) *Let $k, d, t \in \mathbb{N}$, $\lambda > 0$. Let $\boldsymbol{Y} = \lambda v v^\top + \boldsymbol{N}$, where $v \in \mathbb{R}^d$ is a k-sparse unit vector such that $\|v\|_\infty \leqslant 100/\sqrt{k}$ and $\boldsymbol{N}$ is a random matrix with independent (but not necessarily identically distributed) symmetric about zero entries[10] such that for all $i, j \in [d]$, $\mathbb{P}[|\boldsymbol{N}_{ij}| \leqslant 1] \geqslant 0.1$.*

There exists an absolute constant $C > 1$, such that if $k \geqslant Ct \log d$,

$$t \geqslant C \cdot \log\left(2 + d/k^2\right),$$

and

$$\lambda \geqslant k \,,$$

then there exists an algorithm that, given $\boldsymbol{Y}$, k, t and λ, in time $d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v \rangle| \geqslant 0.99 \,.$$

Moreover, we get the same guarantees if we are given only the upper triangle of $\boldsymbol{Y}$, i.e. the entries $\boldsymbol{Y}_{ij}$ such that $i < j$.

If $k = \Theta\left(\sqrt{d}\right)$, this algorithm runs in polynomial time as long as $\lambda \geqslant \varepsilon\sqrt{d}$ for (arbitrary) constant ε. It is the first known polynomial time algorithm for this model, since the algorithm from

10. That is, $\boldsymbol{N}_{ij}$ and $-\boldsymbol{N}_{ij}$ have the same distribution.

d'Orsi et al. (2022) requires quasi-polynomial time. For example, our algorithm finds an estimator close to v or $-v$ in polynomial time if $k = \sqrt{d}$ and $\lambda = k/100$ when the noise has iid Cauchy entries[11] (with location 0 and scale 1), while prior to this work the fastest known algorithm in this regime required quasi-polynomial time even for standard Gaussian noise.

In the special case of Gaussian noise and $\lambda = k$, this algorithm matches the best known algorithmic guarantees. Note, however, that in the regime $k \leqslant d^{1/2-\Omega(1)}$ this algorithm runs in quasipolynomial time. This is not surprising, since as was observed in d'Orsi et al. (2020), sparse PCA with symmetric noise actually generalizes the planted clique problem.

More precisely, let $\boldsymbol{G} \sim G(d, 1/2, k)$ be a random graph with a planted clique of size k. Let $\boldsymbol{A}$ be the adjacency matrix of $\boldsymbol{G}$. Let J be the matrix with all entries equal to 1 and let $\boldsymbol{C} = 2\boldsymbol{A} - J$. Note that the upper triangle of $\boldsymbol{C}$ coincides with the upper triangle of $k \cdot vv^\top + \boldsymbol{\eta}$, where $\sqrt{k} \cdot v$ is the indicator vector of the vertices of the clique (so it is k-sparse), and $\boldsymbol{\eta}$ is the noise whose entries that correspond to the vertices of the clique are zero, and other entries are iid uniform over $\{\pm 1\}$.

The algorithm from theorem 5 solves the planted clique problem in time $n^{O\left(\log\left(2+n/k^2\right)\right)}$, which matches the best known algorithmic guarantees for the planted clique[12]. Moreover, for some $k = n^{\Omega(1)}$ it is conjectured to be impossible to solve it in time $n^{o(\log n)}$ (see Manurangsi et al. (2021) for more details). Note that our algorithm achieves best known algorithmic guarantees for both sparse PCA in the Wigner model and the planted clique problem[13].

2. Techniques

The idea of our approach is similar to the well-known technique of reducing the constant in the planted clique problem. Recall that an instance of the planted clique problem is a random graph $\boldsymbol{G}$ sampled according to the following distribution: First, a graph is sampled from Erdős-Rényi distribution $\mathcal{G}(m, 1/2)$ (i.e. each pair of vertices is chosen independently to be an edge with probability $1/2$), and then a random subset of vertices of size k is chosen (uniformly from the sets of size k and independently from the graph) and the clique corresponding to these vertices is added to the graph. The goal is to find the clique. The problem can be solved in quasi-polynomial time, however, no polynomial time algorithm is known in the regime $k \leqslant o(\sqrt{m})$.

Alon et al. (1998) proposed a spectral algorithm that can be used to find the clique in polynomial time if $k \gtrsim \sqrt{m}$. They also introduced a technique that allows to find the clique in polynomial time if $k \geqslant \varepsilon\sqrt{m}$ for arbitrary constant $\varepsilon > 0$. The idea is to look at every subset T of vertices of size $t \gtrsim \log(1/\varepsilon)$ and consider the subgraph $\boldsymbol{H}(T)$ induced by the vertices of $\boldsymbol{G}$ that are adjacent to T (i.e. adjacent to every vertex of T). This subgraph has approximately $m' = 2^{-t}m \lesssim \varepsilon^2 m$ vertices, and if T was a part of the clique, then the clique is preserved in $\boldsymbol{H}(T)$, and since $k \gtrsim \sqrt{m'}$, we can find a clique applying the spectral algorithm to $\boldsymbol{H}(T)$. The running time of the algorithm is $m^{O(t)}$, so it is polynomial for constant ε.

A similar (but technically more challenging) idea can be also used for sparse PCA. Recall that the instance of sparse PCA (in the Wishart model) is $\boldsymbol{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W}$, where $\boldsymbol{u} \sim N(0,1)^n$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n\times d}$ independent of $\boldsymbol{u}$. To illustrate the idea, we assume that v is flat, i.e. its nonzero entries are $\pm 1/\sqrt{k}$. Instead of the adjacency matrix of the graph, we have the empirical covariance $\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}$. For simplicity, let us ignore cross terms and

11. Cauchy noise is very heavy-tailed, the entries do not even have a finite first moment.
12. Up to a constant factor in the degree.
13. Assuming $\lambda = k$ for sparse PCA.

assume that

$$\tfrac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y} \approx \frac{\|\boldsymbol{u}\|^2}{n}\beta vv^\top + \tfrac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}\,.$$

Since $\boldsymbol{u} \sim N(0, \mathrm{Id})$, $\|\boldsymbol{u}\|^2 \approx n$. So we assume that we are given $\beta vv^\top + \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}$, and the goal is to recover v. Similar to the planted clique, if $\beta \gtrsim \sqrt{d/n}$, there is a spectral algorithm for this problem (that computes the top eigenvector of the empirical covariance).

Suppose that $n \geqslant d$, $k \approx \sqrt{d}$ and $\beta = \varepsilon\sqrt{d/n} \approx \varepsilon k/\sqrt{n}$ for some constant $\varepsilon > 0$. We can look at each subset T of entries of size $t < k$ and try to find a (principal) submatrix of $\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}$ that would be an analogue of the graph $\boldsymbol{H}(T)$ from the algorithm for the planted clique. One option is to say that an entry i is "adjacent" to T if the sum of the elements in the i-th row of $\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}$ is large. However, since v has both positive and negative entries, the sum can be small even if here were no noise and $T \subset \mathrm{supp}(v)$. Hence we also need to take the signs of the entries of v into account.

Let $\mathcal{S}_t$ be the set of all t-sparse vectors with entries from $\{0, \pm 1\}$. Let us call $s \in \mathcal{S}_t$ *correct* if $\mathrm{supp}(s) \subset \mathrm{supp}(v)$ and for all nonzero s_i, $\mathrm{sign}(s_i) = \mathrm{sign}(v_i)$. If $s \in \mathcal{S}_t$ is correct, then

$$\left|\sum_{j\in\mathrm{supp}(s)} \beta v_i v_j s_j\right| = \beta\Big|v_i v^\top s\Big| = \beta t/k\,.$$

For $s \in \mathcal{S}$ let us call an entry $i \in [d]$ *adjacent* to s if either $\big|\big(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}s\big)_i\big| \geqslant \beta t/(2k)$ or $i \in \mathrm{supp}(s)$, and let $\boldsymbol{H}(s)$ be a principal submatrix induced by indices adjacent to s. The size of $\boldsymbol{H}(s)$ is close to pd, where p is the probability that $\big|\big(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}s\big)_i\big|$ is greater than $\beta t/(2k)$. We need to count $i \notin \mathrm{supp}(s)$ adjacent to s. The vector $\boldsymbol{W}s$ has distribution $N(0, t\cdot\mathrm{Id})$, hence $\|\boldsymbol{W}s\| \approx \sqrt{tn}$. Since $i \notin \mathrm{supp}(s)$, the i-th row of $\boldsymbol{W}^\top$ is independent of $\boldsymbol{W}s$, and the distribution of $\big(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}s\big)_i$ is close to $N(0, t/n)$. By the tail bound for the Gaussian distribution,

$$\mathbb{P}\Big[\Big|\Big(\tfrac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}s\Big)_i\Big| \geqslant x\sqrt{t/n}\Big] \leqslant \exp\big(-x^2/2\big)\,.$$

In our case, $x = \frac{\beta\sqrt{tn}}{2k}$. Hence for

$$\beta \gtrsim \frac{k}{\sqrt{tn}}\sqrt{\log(td/k^2)}\,,$$

we get $x \gtrsim \sqrt{\log(td/k^2)}$ and $p = \exp\big(-x^2/2\big) \lesssim k^2/(td)$. Therefore, $\boldsymbol{H}(s)$ has $d' \lesssim k^2/t$ entries. Moreover, by the same argument, $k' \approx (1-p)k \geqslant 0.999k$ entries of v are adjacent to correct s, so the signal part of $\boldsymbol{H}(s)$ is close to $\beta vv^\top$. Since for correct s we get $\beta \gtrsim \sqrt{d'/n}$, we can try to use the spectral algorithm to recover the sparse vector from $\boldsymbol{H}(s)$.

Here we see the difference between planted clique and sparse PCA. In the planted clique problem, if we take a subset of the clique, we can easily recover the whole clique from the output of the spectral algorithm and we do not need to consider other sets after that. In sparse PCA, since we do not know β exactly, it might not be easy to understand if the observed s was correct or not from the output of the spectral algorithm.

We use the following observation: if we have computed the list $\boldsymbol{L} = \{\tilde{\boldsymbol{v}}(s)\}$ of the top eigenvectors of $\boldsymbol{H}(s)$ for all $s \in \mathcal{S}_t$, we can compute a vector close to v (or to $-v$) from this list. Indeed, if we erase all but the largest $O(k)$ entries (in absolute value) of the vectors from $\boldsymbol{L}$, we get a new

list $\boldsymbol{L}'$ of $O(k)$-sparse vectors. It turns out that for correct s not only $\tilde{\boldsymbol{v}}(s)$, but also the corresponding $O(k)$-sparse vector $\boldsymbol{v}'(s) \in \boldsymbol{L}'$ is close to v. Moreover, for all $O(k)$-sparse unit vectors x (in particular, for all vectors in $\boldsymbol{L}'$),

$$x\Big(\tfrac{1}{n}\boldsymbol{W}^\top\boldsymbol{W}\Big)x = 1 \pm \tilde{O}\Big(\sqrt{k/n}\Big)\,.$$

Hence we can just compute $\hat{\boldsymbol{v}} \in \operatorname{argmax}_{x\in\boldsymbol{L}'} x\big(\tfrac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}\big)x$, and it is close to v, since

$$x\Big(\tfrac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}\Big)x \approx 1 + \beta\langle v, x\rangle^2 \pm \tilde{O}\Big(\sqrt{k/n}\Big)\,,$$

and for $\beta > \frac{k}{\sqrt{tn}}$, the term $\tilde{O}\Big(\sqrt{k/n}\Big)$ is smaller than β (as long as $t \ll k$).

Note that in the definition of adjacent entries we used β, which might be unknown. But that is not a problem: if we use some value $\beta/2 < \beta' \leqslant \beta$ instead of β, the algorithm still works. Hence we can use all possible candidates from $n^{-O(1)}$ to $n^{O(1)}$ such that one of them differs from β by at most factor of 2, and in the end work not with the list $\boldsymbol{L}$, but with a list of size $O(\log n)\cdot|\boldsymbol{L}|$.

Remark *[Comparison with the Covariance Thresholding analysis from Deshpande and Montanari (2014)] Our algorithm for $t = 1$ has running time $O(nd^2) + \tilde{O}(d^3)$. For $n \geqslant d$, the running time is comparable to the running time of Covariance Thresholding $O(nd^2)$. The guarantees of both algorithms are the same (up to a constant factor). One advantage of our algorithm is that it is much easier to analyze. The crucial difference is that in Covariance Thresholding one has to bound the spectral norm of thresholded Wishart matrix, which requires a sophisticated probabilistic argument. In our algorithm, we only need to bound principal submatrices of the Wishart matrix, and such bounds easily follow from concentration of the spectral norm of $\big(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W} - \mathrm{Id}\big)$ and a union bound argument. Another advantage of our algorithms is that we can get better guarantees than Covariance Thresholding (by increasing t and hence also the running time) and use the same analysis for all t.*

Remark *[Comparison with the algorithms from Ding et al. (2019)] The algorithms from Ding et al. (2019) also use vectors $s \in \mathcal{S}_t$. However, the crucial difference between our approaches is that they work with $s' \in \mathcal{S}_t$ that maximizes $s\big(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}\big)s$. This approach works only if $\beta \gtrsim \frac{k}{\sqrt{t}}\sqrt{\log d}$, since for smaller β the maximizer of $s\big(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{Y}\big)s$ might be completely unrelated to v. For our analysis it is not a problem, since the correct s is only determined in the end from the list $\boldsymbol{L}'$.*

Adversarial Perturbations. Similar approach also works in the presence of adversarial perturbations, that is, if the input is $Y = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W} + E$ such that the columns of E have norm bounded by $b \ll \beta\sqrt{n/k}$. This is interesting, since known algorithms for sparse PCA that use thresholding techniques and are not based on semidefinite programming, like Diagonal Thresholding, Covariance Thresholding, or the algorithms from Ding et al. (2019), do not work in these settings (see d'Orsi et al. (2020) for more details).

As in the non-adversarial case, we can compute the submatrices $H(s)$ that are induced by indices adjacent to s, that is, indices i such that either $\big|\big(\frac{1}{n}Y^\top Ys\big)_i\big| \geqslant \beta t/(2k)$ or $i \in \operatorname{supp}(s)$. Then, instead of computing the top eigenvector of $H(s)$, we compute $\tilde{X}(s) \in \operatorname{argmax}_{X\in\mathcal{P}_k}\langle X, H(s)\rangle$, where

$$\mathcal{P}_k = \Big\{X \in \mathbb{R}^{d\times d} \;\Big|\; X \succeq 0\,, \operatorname{Tr} X = 1\,, \|X\|_1 \leqslant k\Big\}$$

is the feasible region of the basic SDP for sparse PCA. Then, we can compute the list $\boldsymbol{L}$ of top eigenvectors $\tilde{v}(s)$ of $\tilde{X}(s)$, and perform the same procedure as in the non-adversarial case to recover $\hat{v}$ from $\boldsymbol{L}$.

In order to show the correctness, we need to bound all terms of $\frac{1}{n}Y^\top Y s$. In adversarial settings cross terms can be large, and the most problematic term is $\frac{1}{n}E^\top \boldsymbol{W} s$. Rows of $E^\top$ do not have large norm, but $\boldsymbol{W}s$ has large norm $\|\boldsymbol{W}s\| \approx \sqrt{tn}$, and since E can depend on $\boldsymbol{W}$, the entries of $E^\top \boldsymbol{W} s$ can be large. In particular, for each correct s, the adversary can always choose E such that the term $\frac{1}{n}E^\top \boldsymbol{W} s$ is large enough to make $H(s)$ useless for recovering v.

To resolve this issue, we work with some probability distribution over the set of correct s and show that $\frac{1}{n}E^\top \boldsymbol{W} s$ has small expectation with respect to this distribution. In particular, it implies that for each E there exists some s' such that the term $\frac{1}{n}E^\top \boldsymbol{W} s'$ is small[14]. For flat v, we just divide the support of v into $m = k/t$ blocks of size t, and then each block corresponds to some correct $s \in \mathcal{S}_t$. Then it is enough to consider uniform distribution $\mathcal{U}$ over the set $\{s_1, \dots, s_m\}$ of such s. By the concentration of spectral norm of $\boldsymbol{W}$, with high probability

$$\frac{1}{n^2} \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\mathcal{U}} \langle E_i, \boldsymbol{W}\boldsymbol{s}\rangle^2 = \frac{1}{n^2 m}\sum_{j=1}^{m} \langle E_i, \boldsymbol{W} s_j\rangle^2 \leqslant O\left(\frac{b^2 t}{n^2 m}(m+n)\right) \leqslant O\left(\frac{b^2 t^2}{nk}\right) \ll \frac{\beta^2 t^2}{k^2}\,.$$

Hence there exists some $s' \in \{s_1, \dots, s_m\}$ such that

$$\left\| \left(\frac{1}{n} E^\top \boldsymbol{W} s'\right)_{\mathrm{supp}(v)} \right\|^2 \ll \left\| \beta v v^\top s' \right\|^2 .$$

The other terms of $\frac{1}{n}Y^\top Y s$ can be bounded only assuming that s' is correct (so it is not needed to use properties of $\mathcal{U}$ anymore), hence the signal part of $H(s')$ is close to $\beta v v^\top$.

In addition to pd entries $i \in [d] \setminus \mathrm{supp}(v)$ adjacent to s' that appear due to the term $\frac{1}{n}\boldsymbol{W}^\top \boldsymbol{W} s$, there could be some entries adjacent to s' that appear from $\frac{1}{n}\big(E^\top Y + Y^\top E\big)s'$. We show that the number of such entries is at most $\varepsilon^2 \log(1/\varepsilon) d$, where ε is the same as in theorem 2. Assuming our bound[15] on the maximal norm of columns of E, we get $\varepsilon^2 \log(1/\varepsilon) d \lesssim \beta n$, and by standard properties of basic SDP for sparse PCA, the top eigenvector of $\tilde{X}(s')$ is close to v.

To finish the argument, we need to show that we can still compute $\hat{v}$ close to v or $-v$ from $\boldsymbol{L}$ even in the presence of perturbations. It is not hard since our argument depends only on the upper bound on $x\big(\frac{1}{n}Y^\top Y - \mathrm{Id} - \beta v v^\top\big)x$ for all $O(k)$-sparse x. As was shown in d'Orsi et al. (2020), our assumption on the maximal norm of columns of E is enough to obtain the desired upper bound.

The Wigner model and symmetric noise. In the Wigner model the input is n $\boldsymbol{Y} = \lambda v v^\top + \boldsymbol{W}$. The same argument as for the Wishart model works in these settings, and the proof is technically simpler since there are no cross terms and $\boldsymbol{W}$ is easier to analyze than $\frac{1}{n}\boldsymbol{W}^\top \boldsymbol{W}$ that appears in the Wishart model. Our approach for sparse PCA with perturbations also works for Wigner model, and the proof is much easier in this case since the adversary cannot exploit magnitude of the columns of $\boldsymbol{W}$.

The Wigner model with symmetric noise is more challenging. In these settings we assume that λ is known. We cannot work with $\boldsymbol{Y}s$, since the noise is unbounded. So we first threshold the

14. More precisely, this term has small norm, and it is enough for our analysis.
15. This is the reason why our bound on E is worse than the bound from d'Orsi et al. (2020).

entries of $\boldsymbol{Y}$. Concretely, for $h > 0$ and $x \in \mathbb{R}$, let $\tau_h(x)$ be x if $x \in [-h, h]$ and $\operatorname{sign}(x) \cdot h$ otherwise. We apply this transformation for some $h = \Theta(\lambda/k)$ to the entries of $\boldsymbol{Y}$ and get a new matrix $\boldsymbol{T}$. Then we use our approach for matrix $\boldsymbol{T}$ and for all $s \in \mathcal{S}_t$ we compute the submatrices $\boldsymbol{H}(s)$. As long as $k \gtrsim d$, their algorithm applied to $\boldsymbol{Y}$ outputs a matrix that is close to $\lambda vv^\top$ in polynomial time. As in the Gaussian case, the submatrices $\boldsymbol{H}(s)$ have small size, and we can apply their result to every $\boldsymbol{H}(s)$. However, since $\boldsymbol{H}(s)$ depends on $\boldsymbol{Y}$, the noise part of $\boldsymbol{H}(s)$ might not have the same distribution as $\boldsymbol{N}$. Fortunately, the error probability in d'Orsi et al. (2022) is very small, which allows us to use union bound and conclude that for correct s, the output of the algorithm from d'Orsi et al. (2022) on $\boldsymbol{H}(s)$ is close to $\lambda vv^\top$. Moreover, since we know λ, we do not even need to work with the list of candidates in these settings: It is enough to check the norm of the output, and if it is close to λ, the output is close to $\lambda vv^\top$, and we can recover v from it (up to sign).

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 815464).

I would like to thank David Steurer for helpful conversations. Thanks also Lucas Slot for his comments and suggestions, which greatly enhanced the clarity of the paper. Lastly, I would like to thank the anonymous reviewers whose suggestions substantially improved this paper.

References

Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In Howard J. Karloff, editor, *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 25-27 January 1998, San Francisco, California, USA*, pages 594–598. ACM/SIAM, 1998. URL http://dl.acm.org/citation.cfm?id=314613.315014.

Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, 37(5B):2877–2921, 10 2009. doi: 10.1214/08-AOS664. URL https://doi.org/10.1214/08-AOS664.

Gérard Ben Arous, Alexander S. Wein, and Ilias Zadik. Free energy wells and overlap gap property in sparse PCA. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 479–482. PMLR, 2020. URL http://proceedings.mlr.press/v125/ben-arous20a.html.

Boaz Barak, Jonathan A. Kelner, and David Steurer. Rounding sum-of-squares relaxations. In David B. Shmoys, editor, *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 31–40. ACM, 2014. doi: 10.1145/2591796.2591886. URL https://doi.org/10.1145/2591796.2591886.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 1046–1066. JMLR.org, 2013a.

Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse PCA. *CoRR*, abs/1304.0828, 2013b.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, 41(4):1780–1815, 08 2013c. doi: 10.1214/13-AOS1127. URL https://doi.org/10.1214/13-AOS1127.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 48–166. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/brennan18a.html.

Matthew S. Brennan and Guy Bresler. Optimal average-case reductions to sparse PCA: from weak assumptions to strong hardness. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 469–470. PMLR, 2019. URL http://proceedings.mlr.press/v99/brennan19b.html.

Matthew S. Brennan, Guy Bresler, Samuel B. Hopkins, Jerry Li, and Tselil Schramm. Statistical query algorithms and low degree tests are almost equivalent. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, page 774. PMLR, 2021. URL http://proceedings.mlr.press/v134/brennan21a.html.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 41–48, 2004. URL https://proceedings.neurips.cc/paper/2004/hash/8e065119c74efe3a47aec8796964cf8b-Abstract.html.

Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. In *NIPS*, pages 334–342, 2014.

Ilias Diakonikolas and Daniel Kane. Non-gaussian component analysis via lattice basis reduction. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4535–4547. PMLR, 2022. URL https://proceedings.mlr.press/v178/diakonikolas22d.html.

Jingqiu Ding and Yiding Hua. SQ lower bounds for random sparse planted vector problem. *CoRR*, abs/2301.11124, 2023. doi: 10.48550/arXiv.2301.11124. URL https://doi.org/10.48550/arXiv.2301.11124.

Yunzi Ding, Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Subexponential-time algorithms for sparse PCA. *CoRR*, abs/1907.11635, 2019. URL http://arxiv.org/abs/1907.11635.

Yunzi Ding, Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. The average-case time complexity of certifying the restricted isometry property. *IEEE Trans. Inf. Theory*, 67(11):7355–7361, 2021. doi: 10.1109/TIT.2021.3112823. URL https://doi.org/10.1109/TIT.2021.3112823.

Tommaso d'Orsi, Pravesh K. Kothari, Gleb Novikov, and David Steurer. Sparse PCA: algorithms, adversarial perturbations and certificates. In Sandy Irani, editor, *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pages 553–564. IEEE, 2020. doi: 10.1109/FOCS46700.2020.00058. URL https://doi.org/10.1109/FOCS46700.2020.00058.

Tommaso d'Orsi, Rajai Nasser, Gleb Novikov, and David Steurer. Higher degree sum-of-squares relaxations robust against oblivious outliers. *CoRR*, abs/2211.07327, 2022. doi: 10.48550/arXiv.2211.07327. URL https://doi.org/10.48550/arXiv.2211.07327.

Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017. ISSN 00905364. URL http://www.jstor.org/stable/26362895.

Paul Hand and Laurent Demanet. Recovering the sparsest element in a subspace. *Information and Inference*, 3, 10 2013. doi: 10.1093/imaiai/iau007.

Guy Holtzman, Adam Soffer, and Dan Vilenchik. A greedy anytime algorithm for sparse pca. pages 1939–1956, 2020. URL http://learningtheory.org/colt2020/cfp.html.

Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 178–191. ACM, 2016. doi: 10.1145/2897518.2897529. URL https://doi.org/10.1145/2897518.2897529.

Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. doi: 10.1198/jasa.2009.0121. URL https://doi.org/10.1198/jasa.2009.0121. PMID: 20617121.

Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics*, 43(3):1300 – 1322, 2015. doi: 10.1214/15-AOS1310. URL https://doi.org/10.1214/15-AOS1310.

Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *CoRR*, abs/1907.11636, 2019. URL http://arxiv.org/abs/1907.11636.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse PCA. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1612–1620, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/ec5aa0b7846082a2415f0902f0da88f2-Abstract.html.

Pasin Manurangsi, Aviad Rubinstein, and Tselil Schramm. The strongish planted clique hypothesis and its consequences. In James R. Lee, editor, *12th Innovations in Theoretical Computer Science Conference, ITCS 2021, January 6-8, 2021, Virtual Conference*, volume 185 of *LIPIcs*, pages 10:1–10:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi: 10.4230/LIPIcs.ITCS.2021.10. URL https://doi.org/10.4230/LIPIcs.ITCS.2021.10.

Cheng Mao and Alexander S. Wein. Optimal spectral recovery of a planted vector in a subspace. *CoRR*, abs/2105.15081, 2021. URL https://arxiv.org/abs/2105.15081.

Aaron Potechin and Goutham Rajendran. Sub-exponential time sum-of-squares lower bounds for principal components analysis. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=D45iCWZYcff.

Qing Qu, Zhihui Zhu, Xiao Li, Manolis C. Tsakiris, John Wright, and René Vidal. Finding the sparsest vectors in a subspace: Theory, algorithms, and applications. *CoRR*, abs/2001.06970, 2020. URL https://arxiv.org/abs/2001.06970.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Tengyao Wang, Quentin Berthet, and Richard J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896 – 1930, 2016. doi: 10.1214/15-AOS1369. URL https://doi.org/10.1214/15-AOS1369.

Ilias Zadik, Min Jae Song, Alexander S. Wein, and Joan Bruna. Lattice-based methods surpass sum-of-squares in clustering. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 1247–1248. PMLR, 2022. URL https://proceedings.mlr.press/v178/zadik22a.html.

Appendix A. The Wishart Model

Notation. For $m_1, m_2 \in \mathbb{N}$, we use the notation $\mathbb{R}^{m_1 \times m_2}$ for the set of $m_1 \times m_2$ matrices with entries from $\mathbb{R}$. We denote by $N(0,1)^m$ an m-dimensional random vector with iid standard Gaussian entries. Similarly, we denote by $N(0,1)^{m_1 \times m_2}$ an $m_1 \times m_2$ random matrix with iid standard

Gaussian entries. For $m \in \mathbb{N}$, we denote by $[m]$ the set $\{1, 2, \dots, m-1, m\}$. For a vector $v \in \mathbb{R}^m$, we denote by $\|v\|$ its ℓ_2 norm, and for $p \in [1, \infty]$ we denote by $\|v\|_p$ its ℓ_p norm. For a matrix $M \in \mathbb{R}^{m_1 \times m_2}$, we denote by $\|M\|$ its spectral norm, and by $\|M\|_{\mathrm{F}}$ its Frobenius norm. We write $\log$ for the logarithm to the base e.

Recall that the input is an $n \times d$ matrix $\boldsymbol{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W}$, where $\boldsymbol{u} \sim N(0,1)^n$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n \times d}$ independent of $\boldsymbol{u}$. The goal is to compute a unit vector $\hat{\boldsymbol{v}}$ such that $|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 0.99$ with high probability (with respect to the randomness of $\boldsymbol{u}$ and $\boldsymbol{W}$).

First we define vectors $\boldsymbol{z}_s(r)$ that we will use in the algorithm. Let $t \in \mathbb{N}$ be such that $1 \leqslant t \leqslant k$ and let $\mathcal{S}_t$ be the set of all d-dimensional vectors whose entries are in $-\infty, \prime, \infty$ that have exactly t nonzero coordinates. For $s \in \mathcal{S}_t$ and $r > 0$ let $\boldsymbol{z}_s(r)$ be the d-dimensional (random) vector defined as

$$\boldsymbol{z}_{si}(r) = \begin{cases} \mathbf{1}_{[(\boldsymbol{Y}^\top \boldsymbol{Y} s)_i \geqslant r \cdot t \cdot n]} & \text{if } s_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

The following theorem is a restatement of theorem 1.

Theorem 6 *Let $n, d, k, t \in \mathbb{N}$, $\beta > 0$, $0 < \delta < 0.1$. Let $\boldsymbol{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W}$, where $\boldsymbol{u} \sim N(0,1)^n$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n \times d}$ independent of $\boldsymbol{u}$.*

Suppose that $n \gtrsim k + \frac{t \ln^2 d}{\delta^4}$, $k \gtrsim \frac{t \ln d}{\delta^2}$ and

$$\beta \gtrsim \frac{k}{\delta^2 \sqrt{tn}} \sqrt{\ln\left(2 + \frac{td}{k^2} \cdot \left(1 + \frac{d}{n}\right)\right)}\,.$$

Then there exists an algorithm that, given $\boldsymbol{Y}$, k, t and δ, in time $n \cdot d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 1 - \delta\,.$$

Lemma 7 *For all $s \in \mathcal{S}_t$ such that $s_i = 0$ and for all $\tau > 0$,*

$$\mathbb{P}\Big[\Big(\boldsymbol{W}^\top \boldsymbol{Y} s\Big)_i \geqslant \tau \cdot \Big(\|\boldsymbol{W}s\| + |\langle v, s\rangle| \cdot \sqrt{\beta} \cdot \|\boldsymbol{u}\|\Big) \;\Big|\; \|\boldsymbol{u}\|, \|\boldsymbol{W}s\|\Big] \leqslant \exp\big(-\tau^2/2\big)\,.$$

Proof *Denote $\boldsymbol{g} = \boldsymbol{W}s \sim N(0, t \cdot \mathrm{Id})$.*

$$\Big(\boldsymbol{W}^\top \boldsymbol{Y} s\Big)_i = \langle \boldsymbol{W}_i, \boldsymbol{g}\rangle + \sqrt{\beta} \cdot \langle v, s\rangle \cdot \langle \boldsymbol{W}_i, \boldsymbol{u}\rangle\,.$$

Since $\boldsymbol{W}_i$, $\boldsymbol{g}$ and $\boldsymbol{u}$ are independent, conditional distribution of $\big(\boldsymbol{W}^\top \boldsymbol{Y} s\big)_i$ given $\|\boldsymbol{u}\|$ and $\|\boldsymbol{g}\|$ is $N\Big(0, \big\|\boldsymbol{g} + \sqrt{\beta} \cdot |\langle v, s\rangle| \cdot \boldsymbol{u}\big\|^2\Big)$. The lemma follows from the triangle inequality and the tail bound for Gaussian distribution (fact 38). ∎

Lemma 8 *Let $0 < \delta < 0.1$, $r \leqslant \frac{\beta\delta}{100k}$ and suppose that $n \gtrsim t + \ln^2 d$, $\beta \gtrsim \frac{k}{\delta^2\sqrt{tn}}$ and $k \gtrsim \frac{\ln d}{\delta^2}$.*

Let $s^* \in \operatorname{argmax}_{s\in\mathcal{S}_t}\langle s, v\rangle$. *Then,*[16]

$$\|v \circ \boldsymbol{z}_{s^*}(r)\|^2 = \langle v \circ \boldsymbol{z}_{s^*}(r), v\rangle \geqslant 1-\delta\,,$$

with probability $1 - d^{-10}$.

Proof *To simplify the notation we write* $\boldsymbol{z}_{si}$ *instead of* $\boldsymbol{z}_{si}(r)$. *Let* $\boldsymbol{g} = \boldsymbol{W}s^*$. *By fact 39 with probability at least* $1-\exp(-n/10)$, $n/2 \leqslant \|\boldsymbol{u}\|^2 \leqslant 2n$ *and* $tn/2 \leqslant \|\boldsymbol{g}\|^2 \leqslant 2tn$, *and in this proof we assume that these bounds on* $\|\boldsymbol{u}\|$ *and* $\|\boldsymbol{g}\|$ *are satisfied.*

Let $T = \operatorname{supp}(s^*)$. *Note that it is the set of* t *largest entries of* v *(by absolute value). Since* v *is unit, there are two possibilities: either* $\|v_T\|^2 > 1-\delta$, *or* $\|v_T\|^2 \leqslant 1-\delta$. *If* $\|v_T\|^2 > 1-\delta$, *then the statement is true since* $\|v \circ \boldsymbol{z}_{s^*}\|^2 \geqslant \|v_T\|^2$. *If* $\|v_T\|^2 \leqslant 1-\delta$, *then for every* $j \in T$, $|v_j| \geqslant \sqrt{\delta}/\sqrt{k}$. *Hence*

$$\langle s^*, v\rangle = \|v_T\|_1 \geqslant \sqrt{\delta}\cdot t/\sqrt{k}\,.$$

Let $\mathcal{L} = \left\{ i\in[n] \;\middle|\; |v_i| \geqslant \frac{\sqrt{\delta}}{10\sqrt{k}}\right\}$. *For all* $i \in \mathcal{L}$, $\beta\|\boldsymbol{u}\|^2\langle v, s^*\rangle|v_i| \geqslant 10rtn$. *Note that* $\|v_{\mathcal{L}}\|^2 \geqslant 1-\delta/10$.

Let us write $\boldsymbol{Y}^\top\boldsymbol{Y}s^*$ *as follows*

$$\boldsymbol{Y}^\top\boldsymbol{Y}s^* = \beta\|\boldsymbol{u}\|^2\langle v, s^*\rangle v + \boldsymbol{W}^\top\boldsymbol{W}s^* + \sqrt{\beta}v\boldsymbol{u}^\top\boldsymbol{W}s^* + \sqrt{\beta}\langle v, s^*\rangle\boldsymbol{W}^\top\boldsymbol{u}\,.$$

We will bound each term separately

Consider the term $\boldsymbol{W}^\top\boldsymbol{W}s^* = \boldsymbol{W}^\top\boldsymbol{g}$. *By the Chi-squared tail bound (fact 39), with probability at least* $1-\exp(-\tau/2)$,

$$\left\|\left(\boldsymbol{W}^\top\boldsymbol{g}\right)_{\mathcal{L}\setminus\operatorname{supp}(s^*)}\right\| \leqslant 2\sqrt{tn\cdot(|\mathcal{L}|+\tau)}\,.$$

Since $|\mathcal{L}| \leqslant k$, *with probability at least* $1-\exp(-k)$,

$$\left\|\left(\boldsymbol{W}^\top\boldsymbol{g}\right)_{\mathcal{L}\setminus\operatorname{supp}(s^*)}\right\| \leqslant 10\sqrt{ktn} \leqslant \frac{\delta}{100}\left\|\beta\|\boldsymbol{u}\|^2\langle v, s^*\rangle\cdot v\right\|,$$

where we used $\beta \gtrsim \frac{k}{\delta^2\sqrt{nt}}$.

Consider the term $\sqrt{\beta}\langle v, s^*\rangle\boldsymbol{W}^\top\boldsymbol{u}$. *Since* $\boldsymbol{W}$ *and* $\boldsymbol{u}$ *are independent, by fact 39 with probability at least* $1-\exp(-k/10)$,

$$\left\|\left(\sqrt{\beta}\langle v, s^*\rangle\boldsymbol{W}^\top\boldsymbol{u}\right)_{\mathcal{L}}\right\| \leqslant 2\sqrt{\beta}\|\boldsymbol{u}\|\sqrt{k}\cdot\langle v, s^*\rangle \leqslant \frac{\delta}{100}\left\|\beta\|\boldsymbol{u}\|^2\langle v, s^*\rangle\cdot v\right\|,$$

where we used $\beta n \gtrsim \frac{k\sqrt{n}}{\delta^2\sqrt{t}}$ *and the fact that* $n \geqslant t$.

Consider the term $\sqrt{\beta}v\boldsymbol{u}^\top\boldsymbol{W}s^*$. *With probability at least* $1-\exp(-0.1\sqrt{nt})$,

$$\left\|\sqrt{\beta}v\boldsymbol{u}^\top\boldsymbol{W}s^*\right\| \leqslant 2\sqrt{\beta}\cdot\sqrt{t}\cdot\|\boldsymbol{u}\|\cdot(nt)^{1/4} \leqslant \frac{\delta}{100}\left\|\beta\|\boldsymbol{u}\|^2\langle v, s^*\rangle\cdot v\right\|,$$

where we used the fact that the distribution of $\boldsymbol{u}^\top\boldsymbol{W}s^*$ *given* $\|\boldsymbol{u}\|$ *is* $N(0, t\cdot\|\boldsymbol{u}\|^2)$.

By lemma 26,

$$\|v_{\mathcal{L}} \circ \boldsymbol{z}_{s^*}\|^2 \geqslant (1-\delta/10)\|v_{\mathcal{L}}\|^2 \geqslant 1-\delta\,.$$

■

16. Here and further we denote by $a \circ b$ the entrywise product of vectors $a, b \in \mathbb{R}^d$.

Lemma 9 *Let $r > 0$ and $s \in \mathcal{S}_t$. With probability at least $1 - \delta$, number of entries i such that*

$$\left|\left(\boldsymbol{W}^T Y s\right)_i\right| \geqslant t \cdot r \cdot n$$

is bounded by $pd + 2\sqrt{pd\ln(1/\delta)}$, where $p = \exp\left(-tnr^2/2\right)$.

Proof By lemma 7 and Chernoff bound fact 35, number of such entries is at most $pd + \tau d$ with probability at least $1 - \exp\left(-\frac{\tau^2 d}{2p}\right)$. With $\tau = 2\sqrt{\ln(1/\delta)p/d}$ we get the desired bound. ■

Lemma 10 *For $s \in \mathcal{S}_t$, let $\boldsymbol{N}(s) = \left(\boldsymbol{Y}^\top \boldsymbol{Y} - n \cdot \mathrm{Id} - \beta\|\boldsymbol{u}\|^2 v^\top v\right) \circ \left(\boldsymbol{z}_s(r)\boldsymbol{z}_s^\top(r)\right)$ and let $p = \exp\left(-tnr^2/2\right)$. Suppose that $k \geqslant \ln d$.*

Then for each $s \in \mathcal{S}_t$, with probability $1 - 2d^{-10}$,

$$\|\boldsymbol{N}(s)\| \leqslant 10\sqrt{(n+\beta n)\cdot(pd+k)\ln\left(\frac{ed}{pd+k}\right)} + 10(pd+k)\ln\left(\frac{ed}{pd+k}\right)$$

Proof To simplify the notation we write $\boldsymbol{z}_{si}$ instead of $\boldsymbol{z}_{si}(r)$. By lemma 9, number of nonzero $\boldsymbol{z}_{si}$ is at most $2pd + 2\sqrt{pd\ln(d)} + 2k \leqslant 4pd + 4k$ with probability at least $1 - d^{-10}$. Applying lemma 42, we get the desired bound. ■

Lemma 11 *Let $0 < \delta < 0.1$ and $\frac{\delta\beta}{200k} \leqslant r \leqslant \frac{\delta\beta}{100k}$. Let $p = \exp(-tnr^2/2)$. Suppose that $n \geqslant t\ln^2 d$, $k \geqslant t\ln d$ and*

$$\beta \gtrsim \frac{k}{\delta^2\sqrt{tn}}\sqrt{\ln\left(2 + \frac{td}{k^2}\cdot\left(1+\frac{d}{n}\right)\right)}\,.$$

Then

$$\sqrt{(n+\beta n)\cdot(pd+k)\ln\left(\frac{ed}{pd+k}\right)} + (pd+k)\ln\left(\frac{ed}{pd+k}\right) \leqslant \delta\beta n/10\,.$$

Proof First note that since

$$\ln\left(2 + \frac{td}{k^2} + \frac{td^2}{k^2 n}\right) \geqslant 0.5\cdot\ln\left(2 + \frac{td}{k^2} + \sqrt{\frac{td^2}{k^2 n}}\right),$$

we get

$$\beta \gtrsim \frac{k}{\delta^2\sqrt{tn}}\sqrt{\ln\left(2 + \frac{td}{k^2}\cdot\left(1+\frac{k}{\sqrt{nt}}\right)\right)}\,.$$

Since

$$\sqrt{(n+\beta n)\cdot(pd+k)\ln\left(\frac{ed}{pd+k}\right)} \leqslant \sqrt{(n+\beta n)\cdot pd\ln\left(\frac{e}{p}\right)} + \sqrt{(n+\beta n)\cdot k\ln\left(\frac{ed}{pd+k}\right)},$$

we can bound the term with pd and the term with k separately. Let us bound the first term. Note that

$$\exp(-tnr^2/4) \leqslant \exp\left[-\frac{1}{\delta^2}\ln\left(2+\frac{td}{k^2}\cdot\left(1+\frac{k}{\sqrt{nt}}\right)\right)\right] \leqslant \frac{\delta^2 k}{\sqrt{td+kd\sqrt{t/n}}} \lesssim \min\beta, \sqrt{\beta}\cdot\delta^2\sqrt{n/d}\,.$$

Hence

$$\left(\sqrt{n}+\sqrt{\beta n}\right)\cdot\left(\exp\left(-\frac{tnr^2}{2}\right)tnr^2\sqrt{dn}\right) \leqslant \left(\sqrt{n}+\sqrt{\beta n}\right)\cdot\left(\exp\left(-\frac{tnr^2}{4}\right)\sqrt{dn}\right) \leqslant 0.01\delta^2\beta n\,.$$

Since $t \leqslant \frac{k}{\ln(ed/k)}$ and $n \geqslant k\ln\left(\frac{ed}{k}\right)$, the second term can be bounded as follows

$$\sqrt{(n+\beta n)\cdot k\ln\left(\frac{ed}{pd+k}\right)} \leqslant \sqrt{nk\ln\left(\frac{ed}{k}\right)} + \sqrt{\beta nk\ln\left(\frac{ed}{pd+k}\right)} \leqslant 0.01\delta\beta n\,.$$

Now, let us bound

$$(pd+k)\ln\left(\frac{ed}{pd+k}\right) \leqslant pd\ln(e/p) + k\ln(ed/k)\,.$$

The first term can be bounded as follows:

$$pd\ln(e/p) \leqslant d\cdot\frac{k^2}{td+kd\sqrt{t/n}} \leqslant k\sqrt{n/t} \leqslant 0.01\delta\beta n\,.$$

For the second term,

$$k\ln(ed/k) \leqslant k\sqrt{n/t} \leqslant 0.01\delta\beta n\,.$$

■

Lemma 12 *Let $0 < \delta < 0.1$ and $\frac{\delta\beta}{200k} \leqslant r \leqslant \frac{\delta\beta}{100k}$. For $s \in \mathcal{S}_t$ let $\hat{\boldsymbol{v}}(s)$ be the top*[17] *eigenvector of $\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)\circ\left(\boldsymbol{z}_s(r)\boldsymbol{z}_s^\top(r)\right)$.*

Suppose that $n \gtrsim \frac{t\ln^2 d}{\delta^4}$, $k \gtrsim \frac{t\ln d}{\delta^2}$ and

$$\beta \gtrsim \frac{k}{\delta^2\sqrt{tn}}\sqrt{\ln\left(2+\frac{td}{k^2}\cdot\left(1+\frac{d}{n}\right)\right)}\,.$$

Then there exists $s' \in \mathcal{S}_t$ such that with probability $1-10d^{-10}$,

$$|\langle\hat{\boldsymbol{v}}(s'), v\rangle| \geqslant 1-20\delta\,.$$

Proof Let $s' \in \operatorname{argmax}_{s\in\mathcal{S}_t}\langle s, v\rangle$ and let $\tilde{\boldsymbol{v}} = v\circ\boldsymbol{z}_{s'}(r)$. Then

$$\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)\circ\left(\boldsymbol{z}_{s'}(r)\boldsymbol{z}_{s'}^\top(r)\right) = \beta\|\boldsymbol{u}\|^2\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top + \boldsymbol{N}(s')\,,$$

17. A unit eigenvector that corresponds to the largest eigenvalue

where $\boldsymbol{N}(s')$ is the same as in lemma 10. By lemma 11 and lemma 10, with probability $1-2d^{-10}$,

$$\left\|\boldsymbol{N}(s')\right\| \leqslant \delta\beta n\,.$$

Since with probability at least $1-\exp(-n/10)$, $\|u\|^2 \geqslant n/2$, we can use standard (non-sparse) PCA. Concretely, by lemma 32, $|\langle \hat{\boldsymbol{v}}(s'), \tilde{\boldsymbol{v}}\rangle| \geqslant \|\tilde{v}\|^2 - 2\delta$. By lemma 8, with probability $1-10d^{-10}$, $\|\tilde{v}\|^2 = \langle \tilde{\boldsymbol{v}}, v\rangle \geqslant 1-\delta$. Using fact 33 we get the desired bound. ∎

Proof [Proof of theorem 6] Since $\sqrt{n/2} \leqslant \|\boldsymbol{u}\| \leqslant \sqrt{2n}$ with probability at least $1-\exp(-n/10)$, in this proof we assume that this bound on $\|\boldsymbol{u}\|$ holds.

Let $\tilde{\delta} = \delta/100$. Note that we can apply lemma 12 if $\frac{\tilde{\delta}\beta}{200k} \leqslant r \leqslant \frac{\tilde{\delta}\beta}{100k}$. Since we do not know β, we can create a list of candidates for r of size at most $2\ln n$ (from $r = 1/n$ to $r = 1$).

For all $s \in \mathcal{S}_t$ and for all candidates for r we compute $\hat{\boldsymbol{v}}(s)$ as in lemma 12. By lemma 12, we get a list of vectors $\boldsymbol{L}$ of size $2\ln(n)\cdot|\mathcal{S}_t|$ such that with probability $1-20\ln(n)d^{-10} \geqslant 1-d^{-9}$ there exists $\boldsymbol{v}^* \in \boldsymbol{L}$ such that $|\langle \boldsymbol{v}^*, v\rangle| \geqslant 1-20\tilde{\delta} \geqslant 1-\delta/5$.

By fact 41 and lemma 43, k'-sparse norm of $\boldsymbol{Y}^\top\boldsymbol{Y} - n\mathrm{Id} - \beta\|\boldsymbol{u}\|^2 vv^\top$ is bounded by

$$10\sqrt{nk'\ln(ed/k')} + 10\sqrt{\beta nk'\ln(ed/k')}$$

with probability at least $1-2d^{-9}$. Let us show that if $k' \lesssim \frac{\delta^2\beta^2 n}{(1+\beta)\ln d}$, then this k'-sparse norm is bounded by $\delta\beta n/10$. Indeed, if $\beta \geqslant 1$, then

$$\frac{\delta^2\beta^2 n}{(1+\beta)\ln d} \geqslant \frac{\delta^2\beta n}{2\ln d} \gtrsim k\sqrt{\frac{n}{t\ln^2 d}} \gtrsim k/\delta^2\,,$$

and if $\beta < 1$,

$$\frac{\delta^2\beta^2 n}{(1+\beta)\ln d} \geqslant \frac{\delta^2\beta^2 n}{2\ln^2 d} \gtrsim \frac{k^2}{\delta^2 t\ln d} \gtrsim k/\delta^2\,.$$

Applying lemma 28 with $k' = \lceil 100k/\delta^2\rceil$, we can compute a unit vector $\hat{\boldsymbol{v}}$ such that

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 1-\delta\,.$$

∎

Appendix B. Adversarial Perturbations

The sparse planted vector problem. Let us show that the sparse planted vector problem is a special case of sparse PCA with perturbation matrix $E = -\frac{1}{\|\boldsymbol{u}\|^2}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{W}$. Let $\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{B}$. It follows that

$$\boldsymbol{Y} = \|\boldsymbol{g}_n\|\boldsymbol{R}_n v^\top + \sum_{i=1}^{n-1}\boldsymbol{R}_i\boldsymbol{g}_i^\top\,.$$

Note that $\sum_{i=1}^{n-1}\boldsymbol{R}_i\boldsymbol{g}_i^\top$ has the same distribution as $\left(\mathrm{Id}-\boldsymbol{R}_n\boldsymbol{R}_n^\top\right)\boldsymbol{W}$, where $\boldsymbol{W}\sim N(0,1)^{n\times d}$ is independent of $\boldsymbol{R}$. Hence for Gaussian vector $\boldsymbol{u}$ such that $\frac{1}{\|\boldsymbol{u}\|}\boldsymbol{u} = \boldsymbol{R}_n$, we get

$$\boldsymbol{Y} = \frac{\|\boldsymbol{g}_n\|}{\|\boldsymbol{u}\|}\boldsymbol{u}v^\top + \boldsymbol{W} - \frac{1}{\|\boldsymbol{u}\|^2}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{W}\,.$$

With high probability $\beta := \frac{\|\boldsymbol{g}_n\|^2}{\|\boldsymbol{u}\|^2} = (1+o(1))d/n$. Note that columns of $-\frac{1}{\|\boldsymbol{u}\|^2}\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{W}$ have norm at most $10\sqrt{\log d}$ with high probability. In these settings, ε from theorem 1 is allowed to be as large as $\Omega(1)$, and we get a bound $\|E\|_{1\to 2} \lesssim \sqrt{d/k}$. Hence for $d \gtrsim k \log d$, E is allowed to have columns of norm $10\sqrt{\log d}$.

The Wishart model with perturbations. The following theorem is a restatement of theorem 2.

Theorem 13 *Let $n, d, k, t \in \mathbb{N}$, $\beta > 0$, $0 < \delta < 0.1$. Let*

$$\tilde{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W} + E\,,$$

where $\boldsymbol{u} \sim N(0,1)^n$, $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n\times d}$ independent of $\boldsymbol{u}$ and $E \in \mathbb{R}^{n\times d}$ is a matrix such that

$$\|E\|_{1\to 2} = b \leqslant \varepsilon \cdot \min\sqrt{\beta}, \beta \cdot \sqrt{n/k}\,,$$

where $\|E\|_{1\to 2}$ is the maximal norm of the columns of E and $\varepsilon\sqrt{\ln(1/\varepsilon)} \lesssim \delta^6 \min\Big\{1, \min\big\{\beta, \sqrt{\beta}\big\} \cdot \sqrt{n/d}\Big\}$.

Suppose that $n \gtrsim k + \frac{t\ln^2 d}{\delta^4}$, $k \gtrsim \frac{t\ln d}{\delta^2}$ and

$$\beta \gtrsim \frac{k}{\delta^6\sqrt{tn}}\sqrt{\ln\left(2 + \frac{td}{k^2}\left(1+\frac{d}{n}\right)\right)}\,.$$

Then there exists an algorithm that, given $\boldsymbol{Y}$, k and t, in time $n \cdot d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 1 - \delta\,.$$

Lemma 14 *Let $\mathcal{L} \subset [d]$ be the set of indices of the largest (in absolute value) ℓ entries of v, where $\ell = \min\big\{\lceil \|v\|_1^2/\delta^2\rceil, k\big\}$. Suppose that $\beta \gtrsim \frac{k}{\delta^6\sqrt{nt}}$ and $n \gtrsim k + t\ln^2 d$, $k \gtrsim \ln d$.*

Then with probability $1 - d^{-10}$ there exists $s' \in \mathcal{S}_t$ such that either

$$\left\|\left(\left(\tilde{Y}^\top\tilde{Y} - \beta\|\boldsymbol{u}\|^2 vv^\top\right)s'\right)_{\mathcal{L}\setminus\mathrm{supp}(s')}\right\| \leqslant 10\delta \cdot \left\|\beta\|\boldsymbol{u}\|^2 vv^\top s'\right\|,$$

or

$$\left\|v_{\mathrm{supp}(s')}\right\| \geqslant 1 - \delta.$$

Moreover,

$$\left\|\left(E^\top\boldsymbol{Y} + \boldsymbol{Y}^\top E\right)s'\right\| \leqslant 10^4 \cdot \frac{\varepsilon}{\delta^2} \cdot \sqrt{d} \cdot \frac{\beta n t}{\sqrt{k}\cdot\|v\|_1}\,.$$

Proof First note that

$$\|v_{\mathcal{L}}\|_1 \geqslant \|v\|_1 - \delta \geqslant (1-\delta)\|v\|_1\,.$$

and

$$\|v_{\mathcal{L}}\| \geqslant \|v\| - \|v_{[d]\setminus\mathcal{L}}\|_1 \geqslant \|v\| - \delta \geqslant (1-\delta)\|v\|\,.$$

Also note that for all $i \in \mathcal{L}$, $|v_i| \geqslant \frac{\delta}{\|v\|_1}$.

We will define a distribution over $s \in \mathcal{S}_t$, and then we show via probabilistic method that s' with desired properties exists by bounding terms of $\tilde{Y}^\top \tilde{Y} s$ one by one. Also, with probability at least $1 - \exp(n/2)$, $2\sqrt{n} \geqslant \|u\| \geqslant \sqrt{n}/2$ and in the proof we will always assume that $2\sqrt{n} \geqslant \|u\| \geqslant \sqrt{n/2}$.

Consider the partition of $\mathcal{L}$ into $m = \lceil |\mathcal{L}|/t \rceil$ disjoint blocks $b_1, \dots, b_m$ of size[18] t. Each block b_j corresponds to $s(j) \in \mathcal{S}_t$ such that $s(j)_i = 0$ for $i \notin b_j$ and $s(j)_i = \operatorname{sign}(v_i)$ for $i \in b_j \cap \operatorname{supp}(v)$.

If $|\mathcal{L}| \leqslant t$, then $\left\|v_{\operatorname{supp}(s(1))}\right\| \geqslant \|v_{\mathcal{L}}\| \geqslant 1 - \delta$.

If $|\mathcal{L}| > t$, consider the uniform distribution $\mathcal{U}$ over $s(j)$. We get

$$\mathop{\mathbb{E}}_{\boldsymbol{s} \sim \mathcal{U}} \langle v, \boldsymbol{s} \rangle = \frac{1}{m} \sum_{j=1}^{m} \|v \circ s(j)\|_1 = \frac{1}{m} \|v\|_1 \geqslant \frac{t}{2\mathcal{L}} \|v\|_1 \geqslant \frac{\delta^2 t}{2\|v\|_1} \geqslant \frac{\delta^2 t}{2\sqrt{k}} \,.$$

Moreover,

$$\mathop{\mathbb{E}}_{\boldsymbol{s} \sim \mathcal{U}} \langle v, \boldsymbol{s} \rangle = \frac{1}{m} \|v\|_1 \leqslant \frac{t}{\mathcal{L}} \|v\|_1 \leqslant \frac{t}{\|v\|_1} \,.$$

Let us write $\tilde{Y}^\top \tilde{Y}$ as follows:

$$\tilde{Y}^\top \tilde{Y} = \beta \|\boldsymbol{u}\|^2 v v^\top + \boldsymbol{W}^\top \boldsymbol{W} + \sqrt{\beta} v \boldsymbol{u}^\top \boldsymbol{W} + \sqrt{\beta} \boldsymbol{W}^\top \boldsymbol{u} v^\top + E^\top E + \sqrt{\beta} v \boldsymbol{u}^\top E + \sqrt{\beta} E^\top \boldsymbol{u} v^\top + \boldsymbol{W}^\top E + E^\top \boldsymbol{W} \,.$$

We will bound each term separately.

Consider the term $E\boldsymbol{W}^\top$.

$$\mathop{\mathbb{E}}_{\boldsymbol{s} \sim \mathcal{U}} |\langle E_i, \boldsymbol{W} \boldsymbol{s} \rangle|^2 = \frac{1}{m} \sum_{j=1}^{m} |\langle E_i, \boldsymbol{W} s(j) \rangle|^2$$

Since for different j_1 and j_2, $s(j_1)$ and $s(j_2)$ have disjoint supports, $\boldsymbol{W} s(j) \sim N(0, t)^n$ are independent. By the concentration of spectral norm of Gaussian matrices (fact 40),

$$\sum_{j=1}^{m} \langle E_i, \boldsymbol{W} s(j) \rangle^2 \leqslant 4t(m+n) \cdot \|E_i\|^2 \leqslant 10tnb^2$$

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

$$\mathop{\mathbb{E}}_{\boldsymbol{s} \sim \mathcal{U}} |\langle E_i, \boldsymbol{W} \boldsymbol{s} \rangle|^2 \leqslant \frac{10tb^2 n}{m} \,.$$

Therefore, with probability at least $1 - d\exp(-n)$,

$$\mathop{\mathbb{E}}_{\boldsymbol{s} \sim \mathcal{U}} \sum_{i \in \mathcal{L}} |\langle E_i, \boldsymbol{W} \boldsymbol{s} \rangle|^2 \leqslant |\mathcal{L}| \frac{10tb^2 n}{m}$$

and

$$\mathop{\mathbb{E}}_{\boldsymbol{s} \sim \mathcal{U}} \sum_{i \in [d]} |\langle E_i, \boldsymbol{W} \boldsymbol{s} \rangle|^2 \leqslant d \frac{10tb^2 n}{m}$$

18. If the last block has smaller size, we can add arbitrary entries to it.

Hence with probability at least $1 - d\exp(-n/2)$ there exists $s' = s(j)$ such that

$$\frac{10t}{\|v\|_1} \geqslant \langle v, s(j)\rangle \geqslant \frac{\delta^2 t}{10\|v\|_1} \geqslant \frac{\delta^2 t}{10\sqrt{k}}\,,$$

and

$$\sum_{i\in\mathcal{L}} |\langle E_i, \boldsymbol{W}s(j)\rangle|^2 \leqslant |\mathcal{L}|\frac{40tb^2n}{m} \leqslant 40t^2\varepsilon^2\beta^2n^2/k \leqslant \left(\frac{100\varepsilon}{\delta^2}\right)^2 \left\|\beta\|\boldsymbol{u}\|^2 vv^\top s(j)\right\|^2 \leqslant \delta^2\left\|\beta\|\boldsymbol{u}\|^2 vv^\top s(j)\right\|^2,$$

where we used $|\mathcal{L}| \leqslant mt$, and

$$\sum_{i\in[d]} \left|\langle E_i, \boldsymbol{W}\boldsymbol{s}'\rangle\right|^2 \leqslant \left(\frac{100\varepsilon}{\delta^2}\right)^2 \cdot \frac{d}{k} \cdot \left\|\beta\|\boldsymbol{u}\|^2 vv^\top s(j)\right\|^2.$$

Consider the term $\boldsymbol{W}^\top E$. By fact 40, with probability at least $1 - \exp(-n/2)$,

$$\left\|\boldsymbol{W}^T E s(j)\right\| \leqslant tb\left\|\boldsymbol{W}^\top\right\| \leqslant 10tb\sqrt{n} \leqslant 10t\varepsilon\beta n/\sqrt{k} \leqslant \frac{400\varepsilon}{\delta^2}\left\|\beta\|\boldsymbol{u}\|^2\langle v, s(j)\rangle \cdot v\right\| \leqslant \delta\left\|\beta\|\boldsymbol{u}\|^2 vv^\top s(j)\right\|.$$

Consider the term $\sqrt{\beta}E^\top \boldsymbol{u}v^\top$:

$$\left\|\left(\sqrt{\beta}E^\top \boldsymbol{u}v^\top s(j)\right)_{\mathcal{L}\setminus\mathrm{supp}(s')}\right\| \leqslant \sqrt{k}\left\|\sqrt{\beta}E^\top \boldsymbol{u}v^\top s(j)\right\|_\infty \leqslant \sqrt{k\beta}\cdot b\cdot\|\boldsymbol{u}\|\cdot\langle v, s(j)\rangle \leqslant 10\varepsilon\left\|\beta\|\boldsymbol{u}\|^2\langle v, s(j)\rangle \cdot v\right\|.$$

Consider the term $\sqrt{\beta}v\boldsymbol{u}^\top E$:

$$\left\|\sqrt{\beta}v\boldsymbol{u}^\top E s(j)\right\| \leqslant \sqrt{\beta}tb\|\boldsymbol{u}\| \leqslant \frac{\varepsilon}{\delta^2}\cdot\frac{\delta^2 t}{\sqrt{k}}\cdot\beta\sqrt{n}\|\boldsymbol{u}\| \leqslant \frac{10\varepsilon}{\delta^2}\left\|\beta\|\boldsymbol{u}\|^2\langle v, s(j)\rangle \cdot v\right\|.$$

Consider the term $E^\top E$:

$$\left\|\left(E^\top E s(j)\right)_{\mathcal{L}\setminus\mathrm{supp}(s')}\right\| \leqslant \sqrt{k}\cdot\left\|E^\top E s(j)\right\|_\infty \leqslant \sqrt{k}\cdot b^2 t \leqslant \sqrt{k}\cdot\frac{\varepsilon^2}{\delta^2}\cdot\frac{\delta^2 t}{k}\beta n \leqslant 10\varepsilon\left\|\beta\|\boldsymbol{u}\|^2\langle v, s(j)\rangle \cdot v\right\|.$$

Moreover, note that from the bounds above we also get

$$\left\|\left(E^\top \boldsymbol{Y} + \boldsymbol{Y}^\top E\right)s(j)\right\| \leqslant 1000\cdot\frac{\varepsilon}{\delta^2}\cdot\sqrt{\frac{d}{k}}\cdot\left\|\beta\|\boldsymbol{u}\|^2 vv^\top s(j)\right\| \leqslant 10^4\frac{\varepsilon}{\delta^2}\cdot\sqrt{d}\cdot\frac{\beta nt}{\sqrt{k}\cdot\|v\|_1}\,.$$

Consider the term $\boldsymbol{W}^\top\boldsymbol{W}$. By the Chi-squared tail bound (fact 39), with probability at least $1 - \exp(-\tau/2)$,

$$\left\|\left(\boldsymbol{W}^\top\boldsymbol{W}s(j)\right)_{\mathcal{L}\setminus\mathrm{supp}(s(j))}\right\| \leqslant 2\sqrt{tn\cdot(|\mathcal{L}| + \tau)}\,.$$

Since $|\mathcal{L}| \leqslant k$, with probability at least $1 - \exp(-k)$,

$$\left\|\left(\boldsymbol{W}^\top\boldsymbol{W}s(j)\right)_{\mathcal{L}\setminus\mathrm{supp}(s(j))}\right\| \leqslant 10\sqrt{ktn} \leqslant \delta\left\|\beta\|\boldsymbol{u}\|^2\langle v, s(j)\rangle \cdot v\right\|,$$

where we used $\beta \gtrsim \frac{k}{\delta^3\sqrt{nt}}$.

Consider the term $\sqrt{\beta}\boldsymbol{W}^\top \boldsymbol{u} v^\top$. Since $\boldsymbol{W}$ and $\boldsymbol{u}$ are independent, with probability at least $1-\exp(-k/2)$,

$$\left\| \left(\sqrt{\beta}\boldsymbol{W}^\top \boldsymbol{u} v^\top s(j) \right)_{\mathcal{L}} \right\| \leqslant 2\sqrt{\beta}\|\boldsymbol{u}\|\sqrt{k} \cdot \langle v, s(j)\rangle \leqslant \delta \left\| \beta \|\boldsymbol{u}\|^2 \langle v, s(j)\rangle \cdot v \right\|,$$

where we used $\beta n \gtrsim \frac{k\sqrt{n}}{\delta^2\sqrt{t}}$ and the fact that $n \geqslant t$.

Consider the term $\sqrt{\beta}\boldsymbol{W}^\top \boldsymbol{u} v^\top$. With probability at least $1-\exp(-0.1\sqrt{nt})$,

$$\left\| \sqrt{\beta} v \boldsymbol{u}^\top \boldsymbol{W} s(j) \right\| \leqslant 2\sqrt{\beta} \cdot \sqrt{t} \cdot \|\boldsymbol{u}\| \cdot (nt)^{1/4} \leqslant \delta \left\| \beta \|\boldsymbol{u}\|^2 \langle v, s(j)\rangle \cdot v \right\|,$$

where we used the fact that the distribution of $\boldsymbol{u}^\top \boldsymbol{W} s(j)$ given $\|\boldsymbol{u}\|$ is $\sim N(0, t \cdot \|u\|^2)$. ■

For $s \in \mathcal{S}_t$ let $z_s(r)$ be the n-dimensional (random) vector defined as

$$z_{si}(r) = \begin{cases} \mathbf{1}_{\left[(\tilde{Y}^\top \tilde{Y} s)_i \geqslant r \cdot t \cdot n\right]} & \text{if } s_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

Lemma 15 *Suppose that $k \gtrsim t \ln d$ and*

$$\beta \gtrsim \frac{k}{\delta^2\sqrt{tn}} \sqrt{\ln\left(2 + \frac{td}{k^2} \cdot \left(1 + \frac{d}{n}\right)\right)}.$$

Let r be such that

$$\frac{\beta\delta}{1000\|v\|_1^2} \leqslant r \leqslant \frac{\beta\delta}{500\|v\|_1^2}.$$

Let $s \in \mathcal{S}_t$ and

$$N(s) = \left(\tilde{Y}^\top \tilde{Y} - n\mathrm{Id} - \beta\|\boldsymbol{u}\|^2 v v^\top\right) \circ \left(z_s(r) z_s^\top(r)\right).$$

Let

$$\mathcal{P}_k = \left\{ X \in \mathbb{R}^{d\times d} \;\middle|\; X \succeq 0\,, \operatorname{Tr} X = 1\,, \|X\|_1 \leqslant k \right\}.$$

Then with probability $1 - 10 \cdot d^{-10}$, for all $X \in \mathcal{P}_k$,

$$|\langle X, N\rangle| \leqslant \delta\beta n\,.$$

Proof To simplify the notation we write z instead of $z_s(r)$ and N instead of $N(s)$. Let

$$N_E = \left(\tilde{Y}^\top \tilde{Y} - \left(\tilde{Y} - E\right)^\top \left(\tilde{Y} - E\right)\right) \circ \left(zz^\top\right).$$

Note that

$$|\langle X, N\rangle| \leqslant \|N - N_E\| + |\langle X, N_E\rangle|\,.$$

By lemma 34,

$$|\langle X, N_E\rangle| \leqslant b^2 k + 2b\sqrt{k\left\|\left(\tilde{Y} - E\right)^\top \left(\tilde{Y} - E\right) \circ (zz^\top)\right\|}.$$

The first term can be bounded as follows:

$$b^2 k \leqslant \varepsilon\beta^2 n/k \leqslant \delta\beta\sqrt{n}/100 \leqslant \delta\beta n/100\,.$$

Note that with probability at least $1-\exp(-n)$,

$$\left\|\left(\tilde{Y}-E\right)^\top\left(\tilde{Y}-E\right)\circ\left(zz^\top\right)\right\| \leqslant 10(\beta n+n)\,.$$

Hence for the second term,

$$b\sqrt{k(\beta n+n)} \leqslant \varepsilon\cdot\frac{2\beta}{1+\sqrt{\beta}}n\left(1+\sqrt{\beta}\right) \leqslant \varepsilon\beta n \leqslant \delta\beta n/100\,.$$

Therefore,

$$|\langle X, N_E\rangle| \leqslant \delta\beta n/10\,.$$

By fact 25 and the bound on $\left\|\left(E^\top \boldsymbol{Y}+\boldsymbol{Y}^\top E\right)s'\right\|$ from lemma 14, at most $\tilde{\varepsilon}^2 d = \left(10^4\varepsilon/\delta\right)^2 d$ entries of $\left(E^\top\tilde{Y}+\tilde{Y}^\top E\right)s'$ are larger (in absolute value) than $rtn/2$. By lemma 9, with probability at least $1-d^{-10}$, number of entries i such that

$$\left|\left(\boldsymbol{W}^T Y s'\right)_i\right| \geqslant t\cdot r\cdot n/2$$

is bounded by $pd+10\sqrt{pd\ln d}$, where $p=\exp\left(-tnr^2/8\right)$. Hence number of nonzero entries of z is at most $10\tilde{\varepsilon}^2 d+pd+10\sqrt{pd\ln d}+2k \leqslant 10\tilde{\varepsilon}^2 d+10pd+10k$ with probability at least $1-d^{-10}$.

Therefore, by lemma 42 and lemma 11, with probability at least $1-2d^{-10}$,

$$\|N-N_E\| \leqslant \delta\beta n/10+\sqrt{(n+\beta n)\cdot\tilde{\varepsilon}^2 d\ln(1/\tilde{\varepsilon})}+\tilde{\varepsilon}^2 d\ln(1/\tilde{\varepsilon})\,.$$

Since the second and the third terms are bounded by $\delta\beta n/10$, we get the desired bound. ■

Lemma 16 *Let $0<\delta<0.1$ and let $\frac{\beta\delta^3}{1000\|v\|_1^2} \leqslant r \leqslant \frac{\beta\delta^3}{500\|v\|_1^2}$.*

For $s\in\mathcal{S}_t$ let $\hat{\boldsymbol{v}}(s)$ be the top eigenvector of $X\in\mathcal{P}_k$ that maximizes $\left\langle X, \tilde{Y}\circ\left(\tilde{z}_s\tilde{z}_s^\top\right)\right\rangle$.

Suppose that $n \gtrsim k+t\ln^2 d$, $k\gtrsim t\ln d$ and

$$\beta \gtrsim \frac{k}{\delta^6\sqrt{tn}}\sqrt{\ln\left(2+\frac{td}{k^2}\cdot\left(1+\frac{d}{n}\right)\right)}\,.$$

Then there exists $s'\in\mathcal{S}_t$ such that with probability $1-20d^{-10}$,

$$|\langle\hat{\boldsymbol{v}}(s'), v\rangle| \geqslant 1-\delta/10\,.$$

Proof Let s' be as in lemma 14 and let $\tilde{\boldsymbol{v}} = v\circ z_{s'}(r)$. Then

$$\left(\tilde{Y}^\top\tilde{Y}-n\mathrm{Id}\right)\circ\left(z_{s'}(r)z_{s'}^\top(r)\right) = \beta\|\boldsymbol{u}\|^2\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top+N(s)\,,$$

where $N(s)$ is the same as in lemma 15. By lemma 15, with probability $1 - 10 \cdot d^{-10}$, for all $X \in \mathcal{P}_k$,

$$|\langle X, N\rangle| \leqslant \delta\beta n\,.$$

Consider

$$\hat{\boldsymbol{X}}(s^*) \in \operatorname{argmax}_{X\in\mathcal{P}}\left\langle X, \tilde{Y} \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\rangle.$$

By lemma 32,

$$\left\langle \hat{\boldsymbol{X}}(s^*), vv^\top \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\rangle \geqslant 1 - 6\tilde{\delta}\,.$$

Hence by fact 33,

$$\left\langle \hat{\boldsymbol{X}}(s^*), vv^\top\right\rangle \geqslant 1 - 24\tilde{\delta}\,.$$

By lemma 29, $|\langle \hat{\boldsymbol{v}}(s'), \tilde{\boldsymbol{v}}\rangle| \geqslant 1 - 100\tilde{\delta}$. By lemma 14 and lemma 26, with probability $1 - d^{-10}$,

$$\langle \tilde{\boldsymbol{v}}, v\rangle \geqslant 1 - 100\tilde{\delta}\,.$$

Using fact 33 we get the desired bound. ■

Proof [Proof of theorem 13] Since $\sqrt{n/2} \leqslant \|\boldsymbol{u}\| \leqslant \sqrt{2n}$ with probability at least $1 - \exp(-n/10)$, in this proof we assume that this bound on $\|\boldsymbol{u}\|$ holds. Let $\tilde{\delta} = \delta/1000$. Note that we can apply lemma 15 if $\frac{\beta\tilde{\delta}^3}{1000\|v\|_1^2} \leqslant r \leqslant \frac{\beta\tilde{\delta}^3}{500\|v\|_1^2}$. Since we do not know β and $\|v\|_1$, we can create a list of candidates for r of size at most $2\ln n$ (starting from $r = 1/n$ and finishing at $r = 1$).

For all $s \in \mathcal{S}_t$ and for all candidates for r we compute $\hat{\boldsymbol{v}}(s)$ as in lemma 16. By lemma 16, we get a list of vectors $\boldsymbol{L}$ of size $2\ln(n)\cdot|\mathcal{S}_t|$ such that with probability $1 - 20\ln(n)d^{-10} \geqslant 1 - d^{-9}$ there exists $\boldsymbol{v}^* \in \boldsymbol{L}$ such that $|\langle \boldsymbol{v}^*, v\rangle| \geqslant 1 - \tilde{\delta}/10$.

By fact 41, lemma 43, lemma 34, k'-sparse norm of $\boldsymbol{Y}^\top\boldsymbol{Y} - n\mathrm{Id} - \beta\|\boldsymbol{u}\|^2 vv^\top$ is bounded by

$$10\sqrt{nk'\ln(ed/k')} + 10\sqrt{\beta nk'\ln(ed/k')} + b^2k' + 10b\sqrt{k'(\beta n + n)}$$

with probability at least $1 - 3d^{-9}$. Let us show that if

$$k' \lesssim \min\left\{\delta\beta n/b^2, \frac{\delta^2\beta^2 n}{(1+\beta)\ln d}, \frac{\delta^2\beta^2 n}{(1+\beta)b^2\ln d}\right\},$$

then k' the k'-sparse norm is bounded by $\delta\beta n$. Indeed, if $\beta \geqslant 1$, then

$$\frac{\delta^2\beta^2 n}{(1+\beta)\ln d} \geqslant \frac{\delta^2\beta n}{2\ln d} \gtrsim k\sqrt{\frac{n}{t\ln^2 d}} \gtrsim k/\delta^2\,,$$

and if $\beta < 1$,

$$\frac{\delta^2\beta^2 n}{(1+\beta)\ln d} \geqslant \frac{\delta^2\beta^2 n}{2\ln d} \gtrsim \frac{k^2}{\delta^2 t\ln d} \gtrsim k/\delta^2\,,$$

and by our bound on b,

$$k/\delta^2 \lesssim \min\left\{\delta\beta n/b^2, \frac{\delta^2\beta^2 n}{(1+\beta)b^2\ln d}\right\}.$$

By lemma 28, for $k' = \lceil 100k/\delta^2\rceil$, we can compute a k'-sparse vector unit vector $\hat{\boldsymbol{v}}(k')$ such that $|\langle \hat{\boldsymbol{v}}(k'), v\rangle| \geqslant 1 - \delta$. ■

Appendix C. The Wigner Model

C.1. Classical Settings

In classical settings the input is an $d \times d$ matrix $\boldsymbol{Y} = \lambda v v^\top + \boldsymbol{W}$, where $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{d \times d}$. The goal is to compute a unit vector $\hat{\boldsymbol{v}}$ such that $|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 0.99$ with high probability (with respect to the randomness of $\boldsymbol{W}$).

In this section we prove the following theorem.

Theorem 17 *Let $d, k, t \in \mathbb{N}$, $\lambda > 0$, $0 < \delta < 0.1$. Let $\boldsymbol{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W}$, where $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{d\times d}$.*

Suppose that $k \gtrsim \frac{t \ln d}{\delta^2}$, and

$$\lambda \gtrsim \frac{k}{\delta^2}\sqrt{\frac{\ln(2 + td/k^2)}{t}}\,.$$

Then there exists an algorithm that, given $\boldsymbol{Y}$, k, t and δ, in time $d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 1 - \delta\,.$$

As for Wishart model, we define vectors $\boldsymbol{z}_s(r)$ that we will use in the algorithm. Recall the definition of $\mathcal{S}_t$: for $t \in \mathbb{N}$ such that $1 \leqslant t \leqslant k$ we denote by$\mathcal{S}_t$ the set of all d-dimensional vectors with values in $-\infty, \prime, \infty$ that have exactly t nonzero coordinates. For $s \in \mathcal{S}_t$ let $\boldsymbol{z}_s$ be d-dimensional (random) vectors defined as

$$\boldsymbol{z}_{si}(r) = \begin{cases} \mathbf{1}_{\left[(\boldsymbol{Y}s)_i = \langle s, \boldsymbol{Y}_i\rangle \geqslant r\cdot t\right]} & \text{if } s_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

for some $r > 0$ (here $\boldsymbol{Y}_i$ denotes the i-th row of $\boldsymbol{Y}$).

Lemma 18 *If $i \notin \operatorname{supp}(v)$, then for all $s \in \mathcal{S}_t$ and $r > 0$,*

$$\mathbb{P}[\langle s, \boldsymbol{Y}_i\rangle \geqslant r \cdot t] \leqslant \exp\bigl(-tr^2/2\bigr)\,.$$

Proof *Since $v_i = 0$,*

$$\langle s, \boldsymbol{Y}_i\rangle = \langle s, \boldsymbol{W}_i\rangle \sim N(0, t)\,.$$

The lemma follows from the tail bound for Gaussian distribution (fact 38). ■

Lemma 19 *Let $0 < \delta < 0.1$, $r \leqslant \frac{\lambda\delta}{100k}$ and suppose that $\lambda \gtrsim \frac{k}{\delta^2\sqrt{t}}$ and $k \gtrsim \frac{\ln d}{\delta^2}$.*

Let $s^* \in \operatorname{argmax}_{s\in\mathcal{S}_t}\langle s, v\rangle$. Then, with probability $1 - d^{-10}$

$$\|v \circ \boldsymbol{z}_{s^*}\|^2 = \langle v \circ \boldsymbol{z}_{s^*}, v\rangle \geqslant 1 - \delta\,.$$

Proof *To simplify the notation we write $\boldsymbol{z}_{si}$ instead of $\boldsymbol{z}_{si}(r)$. Let $\mathcal{L}_\delta = \left\{ i \in [d] \;\middle|\; |v_i| \geqslant \frac{\sqrt{\delta}}{10\sqrt{k}}\right\}$. For all $i \in \mathcal{L}$, $\beta\|\boldsymbol{u}\|^2\langle v, s^*\rangle|v_i| \geqslant 10rt$. Note that $\|v_{\mathcal{L}}\|^2 \geqslant 1 - \delta/10$.*

As in the proof of lemma 8, we can assume that

$$\langle s^*, v\rangle = \|v_T\|_1 \geqslant \sqrt{\delta}t/\sqrt{k}\,.$$

We need to bound the norm of $\boldsymbol{W}s^* \sim N(0, t\cdot \mathrm{Id})$ *restricted to the entries of* $\mathcal{L}$*. By fact 39, with probability* $1-\exp(-k/10)$*,*

$$\|(\boldsymbol{W}s^*)_{\mathcal{L}}\| \leqslant 2\sqrt{kt} \leqslant \frac{\delta}{100}\|\lambda\langle v, s^*\rangle v\|\,.$$

By lemma 26,

$$\|v_{\mathcal{L}} \circ \boldsymbol{z}_{s^*}\|^2 \geqslant (1-\delta/10)\|v_{\mathcal{L}}\|^2 \geqslant 1-\delta\,.$$

■

Lemma 20 *Let* $p = \exp\bigl(-tr^2/2\bigr)$*. Then, with probability* $1-d^{-10}$*,*

$$\max_{s\in\mathcal{S}_t}\Bigl\|\boldsymbol{W}\circ\Bigl(\boldsymbol{z}_s(r)\boldsymbol{z}_s^\top(r)\Bigr)\Bigr\| \leqslant 10\sqrt{(pd+k)\cdot\ln\left(\frac{d}{pd+k}\right)}\,.$$

Proof Using the same argument as in the proof of lemma 10, we get that the number of nonzero $\boldsymbol{z}_{si}(r)$ for every $s\in\mathcal{S}_t$ is bounded by $2pd + 2\sqrt{pdt\ln(n/t)} + 2k \leqslant 4pd + 4k$ with probability at least $1-\exp(-pd - t\ln(d/t))$.

By fact 40, an $m\times m$ Gaussian matrix $\boldsymbol{G}$ satisfies

$$\|\boldsymbol{G}\| \leqslant 2\sqrt{m} + \sqrt{\tau}$$

with probability $1-\exp(-\tau/2)$ (for every $\tau > 0$). By union bound over all sets of size at most $4pd+4k$ (corresponding to nonzero rows and columns of $\boldsymbol{W}\circ\bigl(\boldsymbol{z}_s(r)\boldsymbol{z}_s^\top(r)\bigr)$), we get the desired bound. ■

Lemma 21 *Let* $0<\delta<0.1$ *and* $\frac{\delta\lambda}{200k} \leqslant r \leqslant \frac{\delta\lambda}{100k}$*. For* $s\in\mathcal{S}_t$ *let* $\hat{\boldsymbol{v}}(s)$ *be the top eigenvector of* $\boldsymbol{Y}\circ\bigl(\boldsymbol{z}_s(r)\boldsymbol{z}_s^\top(r)\bigr)$*.*

Suppose that $k \gtrsim \frac{t\ln d}{\delta^2}$*, and*

$$\lambda \gtrsim \frac{k}{\delta^2\sqrt{t}}\sqrt{\ln\left(2+\frac{td}{k^2}\right)}\,.$$

Then there exists $s'\in\mathcal{S}_t$ *such that with probability* $1-2d^{-10}$*,*

$$|\langle\hat{\boldsymbol{v}}(s'), v\rangle| \geqslant 1-20\delta\,.$$

Proof Let $s'\in\operatorname{argmax}_{s\in\mathcal{S}_t}\langle s, v\rangle$ and let $\tilde{\boldsymbol{v}} = v\circ\boldsymbol{z}_{s'}(r)$. Then

$$\boldsymbol{Y}\circ\Bigl(\boldsymbol{z}_s(r)\boldsymbol{z}_s^\top(r)\Bigr) = \beta\|\boldsymbol{u}\|^2\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^\top + \boldsymbol{W}\circ\Bigl(\boldsymbol{z}_s(r)\boldsymbol{z}_s^\top(r)\Bigr)\,,$$

By lemma 20 and the same arument as in lemma 11 with $n = d$, with probability at least $1 - d^{-10}$,

$$\left\| \boldsymbol{W} \circ \left(\boldsymbol{z}_s(r) \boldsymbol{z}_s^\top(r) \right) \right\| \leqslant \delta\lambda \,.$$

By lemma 32, $|\langle \hat{\boldsymbol{v}}(u^*), \tilde{\boldsymbol{v}} \rangle| \geqslant 1 - 3\delta$. By lemma 19, with probability $1 - d^{-10}$, $\langle \tilde{\boldsymbol{v}}, v \rangle \geqslant 1 - \delta$. Therefore, by fact 33,

$$\left| \langle \hat{\boldsymbol{v}}(s'), v \rangle \right| \geqslant 1 - 20\delta \,.$$

■

Proof [Proof of theorem 17] For all $s \in \mathcal{S}_t$ we compute $\hat{\boldsymbol{v}}(s)$ as in lemma 21 with $\frac{\delta'\lambda}{200k} \leqslant r \leqslant \frac{\delta'\lambda}{100k}$, where $\delta' = \delta/1000$. Since we do not know λ, we can create a list of candidates for r of size at most $2 \ln d$ (from $r = 1/d$ to $r = 1$).

By lemma 21, we get a list of vectors $\boldsymbol{L}$ of size $2\ln(d)|\mathcal{S}_t|$ such that with probability $1 - d^{-9}$ there exists $\boldsymbol{v}^* \in \boldsymbol{L}$ such that $|\langle \boldsymbol{v}^*, v \rangle| \geqslant 1 - \delta/10$. By fact 41, k'-sparse norm of $\boldsymbol{W}$ is bounded by

$$10\sqrt{k' \ln(ed/k')}$$

with probability at least $1 - d^{-10}$. Hence by lemma 28 with $k' = \lceil 100k/\delta^2 \rceil$, we get the desired estimator. ■

C.2. Adversarial Perturbations

The following theorem is the restatement of theorem 4.

Theorem 22 *Let $d, k, t \in \mathbb{N}$, $\lambda > 0$, $0 < \delta < 0.1$. Let $\tilde{Y} = \sqrt{\beta}\boldsymbol{u}v^\top + \boldsymbol{W} + E$, where $v \in \mathbb{R}^d$ is a k-sparse unit vector, $\boldsymbol{W} \sim N(0,1)^{n \times d}$, and $E \in \mathbb{R}^{d \times d}$ is a matrix with entries*

$$\|E\|_\infty = \varepsilon\lambda/k \lesssim \delta^3\lambda/k \,.$$

Suppose that $k \gtrsim \frac{t \ln d}{\delta^2}$ and

$$\lambda \gtrsim \frac{k}{\delta^2}\sqrt{\frac{\ln(2 + td/k^2)}{t}} \,.$$

Then there exists an algorithm that, given $\tilde{Y}$, k, t and δ, in time $d^{O(t)}$ outputs a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v \rangle| \geqslant 1 - \delta \,.$$

Proof Let $\tilde{\delta} = \delta/1000$ and let $r = \frac{\tilde{\delta}}{100k}$. For $s \in \mathcal{S}_t$ let $\tilde{z}_s$ be n-dimensional (random) vectors defined as

$$\tilde{z}_{si}(r) = \begin{cases} \mathbf{1}_{[\langle s, \tilde{Y}_i \rangle \geqslant r \cdot t]} & \text{if } s_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

Let $\boldsymbol{Y} = \tilde{Y} - E$, and $\boldsymbol{z}_{si}$ be the same as in the non-adversarial case defined for $\boldsymbol{Y}$. Note that

$$\boldsymbol{z}_{si}(r - \varepsilon \cdot \lambda/k) \leqslant \tilde{z}_{si}(r) \leqslant \boldsymbol{z}_{si}(r + \varepsilon \cdot \lambda/k) \,.$$

Let $s^* \in \operatorname{argmax}_{s \in \mathcal{S}_t} \langle s, v \rangle$. By lemma 19, with probability $1 - d^{-10}$,

$$\|v \circ \tilde{z}_{s^*}\|^2 = \langle v \circ \tilde{z}_{s^*}, v \rangle \geqslant \boldsymbol{z}_{si}(r - \varepsilon \cdot \lambda/k) \geqslant 1 - 2\tilde{\delta}\,.$$

By the argument from the proof of lemma 10, with probability $1 - d^{-10}$, number of nonzero entries of $\tilde{z}_s$ is at most $4pd + 4k$ and

$$\left\|\boldsymbol{W} \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\| \leqslant 10\sqrt{(pd+k) \cdot \ln\left(\frac{d}{pd+k}\right)} \lesssim \tilde{\delta}^2\lambda\,.$$

Let $\mathcal{P}_k = \left\{X \in \mathbb{R}^{d\times d} \;\middle|\; X \succeq 0\,, \operatorname{Tr} X = 1\,, \|X\|_1 \leqslant k\right\}$. For all $X \in \mathcal{P}_k$,

$$\left|\left\langle X, \boldsymbol{W} \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\rangle\right| \leqslant \left\|\boldsymbol{W} \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\| \cdot \|X\| \lesssim \tilde{\delta}\lambda\,.$$

and

$$\left|\left\langle X, E \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\rangle\right| \leqslant \left\|E \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\|_\infty \cdot \|X\|_1 \leqslant \tilde{\delta}\lambda\,.$$

Consider

$$\hat{X}(s^*) \in \operatorname{argmax}_{X\in\mathcal{P}}\left\langle X, \tilde{Y} \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\rangle.$$

By lemma 32,

$$\left\langle \hat{X}(s^*), vv^\top \circ \left(\tilde{z}_{s^*}\tilde{z}_{s^*}^\top\right)\right\rangle \geqslant 1 - 6\tilde{\delta}\,.$$

Hence by fact 33,

$$\left\langle \hat{X}(s^*), vv^\top\right\rangle \geqslant 1 - 24\tilde{\delta}\,.$$

Let $\hat{v}(s^*)$ be the top eigenvector of $\hat{X}(s^*)$. By lemma 29,

$$|\langle \hat{v}(s^*), v\rangle| \geqslant 1 - 100\tilde{\delta}\,.$$

By fact 41, k'-sparse norm of $\boldsymbol{W}$ is bounded by

$$10\sqrt{k' \ln(ed/k')}$$

with probability at least $1 - d^{-10}$. And k'-sparse norm of E is bounded by

$$k'\|E\|_\infty \leqslant \varepsilon\lambda\frac{k'}{k}\,.$$

Hence by lemma 28 with $k' = \lceil 100k/\tilde{\delta}^2\rceil$ and the list of $\boldsymbol{X}(s)$ for all $s \in \mathcal{S}_t$, we get the desired estimator. ∎

Appendix D. Heavy-tailed Symmetric Noise

The following theorem is a restatement of theorem 5.

Theorem 23 *Let $k, d, t \in \mathbb{N}$, $\lambda > 0$, $A \geqslant 1$, $0 < \alpha < 1$, $0 < \delta < 0.1$. Let*

$$\boldsymbol{Y} = \lambda v v^\top + \boldsymbol{N}\,,$$

where $v \in \mathbb{R}^d$ is a k-sparse unit vector such that $\|v\|_\infty \leqslant A/\sqrt{k}$ and $\boldsymbol{N} \sim N(0,1)^{d\times d}$ is a random matrix with independent (but not necessarily identically distributed) symmetric about zero entries such that

$$\mathbb{P}[|\boldsymbol{N}_{ij}| \leqslant 1] \geqslant \alpha\,.$$

Suppose that $k \gtrsim \frac{t \ln d}{\delta^4 A^4 \alpha^2}$,

$$t \gtrsim \frac{\ln\left(2 + td/k^2\right)}{\alpha^2 A^4 \delta^6}\,,$$

and

$$\lambda \geqslant k\,.$$

Then there exists an algorithm that, given $\boldsymbol{Y}$, k, t and λ, in time $d^{O(t)}$ finds a unit vector $\hat{\boldsymbol{v}}$ such that with probability $1 - o(1)$ as $d \to \infty$,

$$|\langle \hat{\boldsymbol{v}}, v\rangle| \geqslant 1 - O(\delta)\,.$$

Before proving this theorem, we state here a theorem from d'Orsi et al. (2022).

Theorem 24 (d'Orsi et al. (2022)) *Let $\delta, \alpha \in (0,1)$ and $\zeta \geqslant 0$. Let $\tilde{\Omega} \subseteq \mathbb{R}^m$ be a compact convex set. Let $b, r, \gamma \in \mathbb{R}$ be such that*

$$\max_{X\in\tilde{\Omega}} \|X\|_\infty \leqslant b\,,$$

$$\max_{X\in\tilde{\Omega}} \|X\|_2 \leqslant r\,,$$

and

$$\mathop{\mathbb{E}}_{\boldsymbol{W}\sim N(0,\mathrm{Id})}\left[\sup_{X\in\tilde{\Omega}} \langle X, \boldsymbol{W}\rangle\right] \leqslant \gamma\,.$$

Consider

$$\boldsymbol{Y} = X^* + \boldsymbol{N}\,,$$

where $X^* \in \tilde{\Omega}$ and $\boldsymbol{N}$ is a random m-dimensional vector with independent (but not necessarily identically distributed) symmetric about zero entries satisfying $\mathbb{P}[|\boldsymbol{N}_i| \leqslant \zeta] \geqslant \alpha$.

Then the minimizer $\hat{X} = \operatorname{argmin}_{X\in\tilde{\Omega}} F_h(\boldsymbol{Y} - X)$ of the Huber loss with parameter $h \geqslant 2b + \zeta$ satisfies

$$\left\|\hat{X} - X^*\right\|_2 \leqslant O\left(\sqrt{\frac{h}{\alpha}\left(\gamma + r\sqrt{\log(1/\delta)}\right)}\right)$$

with probability at least $1 - \delta$ over the randomness of $\boldsymbol{N}$.

Using this result, we will prove theorem 23

Proof [Proof of theorem 23] Consider

$$\tau_h(x) := \begin{cases} h & \text{if } x > h \\ x & \text{if } |x| \leqslant h \\ -h & \text{if } x < -h \end{cases}$$

where $h = 3\lambda\|v\|_\infty^2 \leqslant 3\frac{\lambda}{k}A^2$. Let $\boldsymbol{T} = \tau_h(\boldsymbol{Y})$ (i.e. the matrix obtained from Y by applying τ_h to each entry).

For $s \in \mathcal{S}_t$ let $\boldsymbol{z}_s$ be n-dimensional (random) vectors defined as

$$\boldsymbol{z}_{si}(r) = \begin{cases} \mathbf{1}_{[(\boldsymbol{T}s)_i = \langle s, \boldsymbol{T}_i\rangle \geqslant r\cdot t]} & \text{if } s_i = 0 \\ 1 & \text{otherwise} \end{cases}$$

Let $\frac{\lambda\delta^2\alpha}{10k} \leqslant r \leqslant \frac{\lambda\delta^2\alpha}{5k}$.

Let $\mathcal{L} = \left\{ i \in [d] \;\middle|\; |v_i| \geqslant \frac{\delta}{\sqrt{k}} \right\}$. Note that $\|v_{\mathcal{L}}\|^2 \geqslant 1 - \delta^2$ and hence $|\mathcal{L}| \geqslant \frac{k}{2A^2}$.

Let $s^* \in \operatorname{argmax}_{s\in\mathcal{S}_t}\langle s, v\rangle$. By the same argument as in the proof of lemma 8,

$$\langle s^*, v\rangle = \|v_T\|_1 \geqslant \delta t/\sqrt{k}\,.$$

By Chernoff bound, for every $i \in \mathcal{L}$ with probability at least $1 - \exp(-t\alpha/2)$, for at least $\alpha t/10$ of $j \in \operatorname{supp}(s^*(j))$, $|\boldsymbol{N}_{ij}| \leqslant 1$. Let $\mathcal{C}_i$ be the set of such entries. Then for $j \in \mathcal{C}_i$,

$$\tau_h(\boldsymbol{Y}_{ij}) = \lambda v_i v_j + |\boldsymbol{N}_{ij}|\,.$$

By Hoeffding's inequality,

$$\sum_{j\in\mathcal{C}_i} \tau_h(\boldsymbol{Y}_{ij}) \geqslant \lambda\frac{\delta|\mathcal{C}_i|}{\sqrt{k}}v_i - \sqrt{|\mathcal{C}_i|q}$$

with probability at least $1 - \exp(-q/2)$ Note that for all $j \in \operatorname{supp}(j)$,

$$\mathbb{E}\,\tau_h(\boldsymbol{Y}_{ij}) \geqslant 0\,.$$

By Hoeffding's inequality,

$$\sum_{j\in\operatorname{supp}(s^*(j))\setminus\mathcal{C}_i} \tau_h(\boldsymbol{Y}_{ij}) \geqslant -h\sqrt{tq}$$

with probability at least $1 - \exp(-q/2)$. Hence for $i \in \mathcal{L}$, with probability at least $1 - \delta^2/A^2$.

$$\langle s^*, \boldsymbol{T}_i\rangle \geqslant \frac{\lambda\delta^2\alpha t}{2k} - 10h\sqrt{t\log(A/\delta)} \geqslant \frac{\lambda\delta^2\alpha t}{4k} \geqslant r\cdot t\,.$$

Let $\mu = \delta^2/A^2$. By Chernoff bound, with probability at least $1 - \exp(-\mu k/10)$, for at most $2\mu k$ entries $i \in \mathcal{L}$, $\boldsymbol{z}_{s^*i} = 0$. Hence with probability at least $1 - \exp(-\mu|\mathcal{L}|/100)$,

$$\|v_{\mathcal{L}} \circ \boldsymbol{z}_{s^*}\|^2 \geqslant 1 - 2A^2\mu \geqslant 1 - \delta^2\,.$$

By Hoeffding's inequality, for all $i \notin \operatorname{supp}(v)$,

$$|\langle s^*, \boldsymbol{T}_i \rangle| \leqslant h\sqrt{tq}$$

with probability at least $1 - \exp(-q/2)$. Hence

$$|\langle s^*, \boldsymbol{T}_i \rangle| < r \cdot t$$

with probability at least $1 - \exp\left(-\frac{r^2 t}{10h^2}\right) = 1 - p$.

By the same argument as in lemma 10, number of nonzero $\boldsymbol{z}_{s^* i}(r)$ is at most $4pd + 4k$ with probability at least $1 - \exp(-pd - t\ln(d/t))$. For $s \in \mathcal{S}_t$, denote by $\boldsymbol{Z}(s)$ the set of $i \in [d]$ such that $\boldsymbol{z}_{si}(r) = 1$.

For $Q \subset [d]$ let

$$\mathcal{P}_Q = \left\{ X \in \mathbb{R}^{Q \times Q} \;\middle|\; X \succeq 0\,, \operatorname{Tr} X \leqslant \lambda\,, \|X\|_1 \leqslant \lambda k \right\},$$

Note that

$$\gamma(Q) := \mathop{\mathbb{E}}_{\boldsymbol{G} \sim N(0,1)^{Q \times Q}} \left[\sup_{X \in \boldsymbol{\mathcal{P}}_Q} \langle X, \boldsymbol{G} \rangle \right] \leqslant \lambda \cdot \mathop{\mathbb{E}}_{\boldsymbol{G} \sim N(0,1)^{Q \times Q}} \|\boldsymbol{G}\| \leqslant 10\lambda\sqrt{|Q|}\,.$$

Hence by theorem 24 and a union bound over all sets Q of size at most $4pd + 4k$, we get with probability at least $1 - d^{-10}$, for all $s \in \mathcal{S}_t$,

$$\left\| \hat{\boldsymbol{X}}(s) - \lambda \tilde{\boldsymbol{v}}(s) \tilde{\boldsymbol{v}}(s)^\top \right\|_{\mathrm{F}}^2 \leqslant O\left(\frac{h}{\alpha} \cdot \lambda \sqrt{(pd+k) \cdot \ln\left(\frac{d}{pd+k}\right)} \right) \leqslant O\left(\frac{A^2}{\alpha} \cdot \frac{\lambda^2}{k} \cdot \sqrt{(pd+k) \cdot \ln\left(\frac{d}{pd+k}\right)} \right),$$

where $\hat{\boldsymbol{X}}(s)$ is the minimizer of the Huber loss with parameter h over $\mathcal{P}_{\boldsymbol{Z}(s)}$ and $\tilde{\boldsymbol{v}} = v \circ \boldsymbol{z}_s(r)$.

Note that

$$\sqrt{(pd+k) \cdot \ln\left(\frac{d}{pd+k}\right)} \lesssim \sqrt{pd\ln(1/p)} + \sqrt{k \ln d}\,.$$

The second term can be bounded as follows

$$\sqrt{k \ln d} \lesssim \delta^2 k \frac{\alpha}{A^2}\,.$$

Note that

$$r^2 t / h^2 \geqslant \frac{\delta^4 \alpha^2}{A^4} t \gtrsim \ln\left(2 + td/k^2\right).$$

Hence the first term can be bounded as follows

$$\sqrt{pd\ln(1/p)} \lesssim \frac{k}{\sqrt{d}} \cdot \frac{\alpha\delta^2}{A^2} \cdot \sqrt{d} \leqslant \delta^2 k \frac{\alpha}{A^2}\,.$$

Hence for all $s \in \mathcal{S}_t$,

$$\left\| \hat{\boldsymbol{X}}(s) - \lambda \tilde{\boldsymbol{v}}(s) \tilde{\boldsymbol{v}}(s)^\top \right\|_{\mathrm{F}}^2 \leqslant \delta^2 \lambda^2\,.$$

Since $\|v \circ \boldsymbol{z}_{s^*}\|^2 \geqslant 1 - 2\delta^2$,

$$\left\|\hat{\boldsymbol{X}}(s^*)\right\|_{\mathrm{F}} \geqslant 1 - 10\delta\lambda\,.$$

Consider some s' such that $\left\|\hat{\boldsymbol{X}}(s')\right\|_{\mathrm{F}} \geqslant 1 - 10\delta\lambda$. For such s',

$$\langle v \circ \boldsymbol{z}_{s'}(r), v\rangle = \|v \circ \boldsymbol{z}_{s'}(r)\|^2 \geqslant 1 - 100\delta\,.$$

Moreover, since $\left\|\hat{\boldsymbol{X}}(s') - \lambda\tilde{\boldsymbol{v}}(s')\tilde{\boldsymbol{v}}(s')^\top\right\|_{\mathrm{F}} \leqslant \delta\lambda$, the top eigenvector $\hat{\boldsymbol{v}}(s')$ of $\hat{\boldsymbol{X}}(s')$ satisfies

$$|\langle\hat{\boldsymbol{v}}(s'), v\rangle| \geqslant 1 - O(\delta)\,.$$

■

Appendix E. Properties of sparse vectors

This section contain tools used throughout the rest of the paper.

Fact 25 *Let $r, \delta > 0$ and let $x \in \mathbb{R}^m$ such that $\|x\| \leqslant R$. Let $\mathcal{S} = \{i \in [m] \mid |x_i| \geqslant \delta\}$. Then*

$$|\mathcal{S}| \leqslant R^2/\delta^2\,.$$

Proof

$$\delta^2 \cdot |\mathcal{S}| \leqslant \|x_{\mathcal{S}}\|^2 \leqslant \|x\|^2 \leqslant r^2\,.$$

■

Lemma 26 *Let $\delta, \delta' \in (0,1)$. Let $x, y \in \mathbb{R}^m$ such that $\|y\| \leqslant \delta\|x\|$. Let $\mathcal{S} = \{i \in [m] \mid |y_i| \geqslant \delta'|x_i|\}$. Then*

$$\|v_{\mathcal{S}}\| \geqslant \big(1 - \delta/\delta'\big)\|v\|\,.$$

Proof Consider the vector y' such that $y'_i = -x_i$ for all $i \in \mathcal{S}$ and $y'_i = 0$ for all $i \notin \mathcal{S}$. It follows that

$$\|y'\| \leqslant \frac{1}{\delta'}\|y\| \leqslant \frac{\delta}{\delta'}\|x\|\,.$$

Hence

$$\|x_{\mathcal{S}}\| = \|x + y'\| \geqslant \|x\|_2 - \|y'\| \geqslant \big(1 - \delta/\delta'\big)\|x\|\,.$$

■

Lemma 27 *Let $v \in \mathbb{R}^d$ be a k-sparse unit vector, and suppose that for some unit vector $v' \in \mathbb{R}^d$, $|\langle v, v'\rangle| \geqslant 1 - \delta$. For $k' \geqslant k$, let $\mathcal{K}'$ be the set of k' largest (in absolute value) entries of v'. Then*

$$\big|\big\langle v, v'_{\mathcal{K}'}\big\rangle\big| \geqslant 1 - \delta - \sqrt{k/k'}\,.$$

Proof Since $\|v\| = 1$, $\left\|v'_{\mathcal{K}'} - v'\right\|_\infty \leqslant 1/k'$. Hence

$$\left|\left\langle v, v'_{\mathcal{K}'} - v'\right\rangle\right| \leqslant \left\|v'_{\mathcal{K}'} - v'\right\|_\infty \cdot \|v\|_1 \leqslant \sqrt{k/k'}\,.$$

Therefore,

$$\left|\left\langle v, v'_{\mathcal{K}'}\right\rangle\right| \geqslant \left|\left\langle v, v'\right\rangle\right| - \left|\left\langle v, v'_{\mathcal{K}'} - v'\right\rangle\right| \geqslant 1 - \delta - \sqrt{k/k'}\,.$$

■

Lemma 28 *Let $\lambda, \kappa, \delta > 0$ and let $v \in \mathbb{R}^d$ be a k-sparse unit vector. Let $L \subset \mathbb{R}^d$ be a finite set of unit vectors such that*

$$\max_{x \in L} |\langle v, x\rangle| \geqslant 1 - \delta\,.$$

Let $k' \geqslant 2\lceil k/\delta^2\rceil$ and let $N \in \mathbb{R}^{d\times d}$ be a matrix such that for every k'-sparse unit vector $u \in \mathbb{R}^d$

$$u^\top N u \leqslant \kappa\,.$$

Then, there exists an algorithm running in time $O\big(d^2 \cdot L\big)$ that, given $Y = \lambda v v^\top + N$, L, k and δ as input, finds a unit vector $\hat{v}$ such that

$$|\langle v, \hat{v}\rangle| \geqslant 1 - 4\delta - \frac{2\kappa}{\lambda}\,.$$

Proof For each $x \in L$ we can compute a k'-sparse vector $s(x)$ that coincides with x on the top k' largest (in absolute value) entries. Let $x^* \in \operatorname{argmax}_{x\in L}|\langle v, s(x)\rangle|$. By lemma 27,

$$|\langle v, s(x^*)\rangle| \geqslant 1 - 2\delta\,.$$

Hence

$$s(x^*)^\top \Big(\lambda v v^\top + N\Big) s(x^*) \geqslant (1-2\delta)^2\lambda - \kappa \geqslant (1-4\delta)\lambda - \kappa\,.$$

Let

$$\tilde{v} \in \operatorname{argmax}_{s(x)\,, x\in L} s(x)^\top \Big(\lambda v v^\top + N\Big) s(x)\,.$$

Then

$$\tilde{v}^\top \Big(\lambda v v^\top + N\Big) \tilde{v} \geqslant s(x^*)^\top \Big(\lambda v v^\top + N\Big) s(x^*) \geqslant (1-4\delta)\lambda - \kappa\,.$$

Since $\tilde{v}^\top N \tilde{v} \leqslant \kappa$, we get

$$|\langle v, \tilde{v}\rangle| \geqslant |\langle v, \tilde{v}\rangle|^2 \geqslant 1 - 4\delta - 2\kappa/\lambda\,.$$

Hence $\hat{v} = \frac{1}{\|\tilde{v}\|}\tilde{v}$ is the desired estimator. ■

Appendix F. Linear Algebra

Lemma 29 *Let $M \in \mathbb{R}^{d\times d}$, $M \succeq 0$, $\operatorname{Tr} M = 1$ and let $z \in \mathbb{R}^d$ be a unit vector such that $z^{\mathsf{T}} M z \geqslant 1-\varepsilon$. Then the top eigenvector v_1 of M satisfies $\langle v_1, z\rangle^2 \geqslant 1-2\varepsilon$.*

Proof *Write $z = \alpha v_1 + \sqrt{1-\alpha^2} v_\perp$ where $v_\perp$ is a unit vector orthogonal to v_1.*

$$\begin{aligned} z^{\mathsf{T}} M z &= \alpha^2 {v_1}^{\mathsf{T}} M v_1 + \left(1-\alpha^2\right) {v_\perp}^{\mathsf{T}} M v_\perp \\ &= \alpha^2 \left(\lambda_1 - {v_\perp}^{\mathsf{T}} M v_\perp\right) + {v_\perp}^{\mathsf{T}} M v_\perp \\ &\geqslant 1-\varepsilon \end{aligned}$$

As ${v_1}^{\mathsf{T}} M v_1 \geqslant z^{\mathsf{T}} M z$ and ${v_\perp}^{\mathsf{T}} M v_\perp \leqslant 1 - {v_1}^{\mathsf{T}} M v_1 \leqslant 1 - z^{\mathsf{T}} M z \leqslant \varepsilon$, rearranging

$$\alpha^2 \geqslant \frac{1-\varepsilon - {v_\perp}^{\mathsf{T}} M v_\perp}{\lambda_1 - {v_\perp}^{\mathsf{T}} M v_\perp} \geqslant 1-2\varepsilon.$$

■

Fact 30 *Let $A, B \in \mathbb{R}^{d\times d}$, $A, B \succeq 0$. Then $\langle A, B\rangle \geqslant 0$.*

Fact 31 *Let $X \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix. Then for all $A \in \mathbb{R}^{d\times d}$,*

$$|\langle A, X\rangle| \leqslant \|A\| \cdot \operatorname{Tr} X\,.$$

Lemma 32 *Let $\Omega \subset \mathbb{R}^m$. Let $Y = S + N$, where $S \in \Omega$ and $N \in \mathbb{R}^m$ statisfies*

$$\sup_{X\in\Omega} |\langle X, N\rangle| \leqslant \delta\,.$$

Then $\hat{X} \in \operatorname{argmax}_{X\in\Omega}\langle X, Y\rangle$ satisfies

$$\left\langle \hat{X}, S\right\rangle \geqslant \|S\|^2 - 2\delta\,.$$

Proof

$$\left\langle \hat{X}, S\right\rangle = \left\langle \hat{X}, Y\right\rangle - \left\langle \hat{X}, N\right\rangle \geqslant \left\langle \hat{X}, Y\right\rangle - \delta \geqslant \langle S, Y\rangle - \delta = \langle S, S\rangle + \langle S, N\rangle - \delta \geqslant \|S\|^2 - 2\delta\,.$$

■

Fact 33 *Let $a, b, c \in \mathbb{R}^m$ such that $\|a\| = \|c\| = 1$ and $\|b\| \leqslant 1$. Suppose that $\langle a, b\rangle \geqslant 1-\delta$ and $\langle c, b\rangle \geqslant 1-\delta$. Then $\langle a, c\rangle \geqslant 1-4\delta$.*

Proof Note that $\|a-b\|^2 \leqslant 2 - 2\langle a, b\rangle \leqslant 2\delta$ and similarly $\|c-b\|^2 \leqslant 2\delta$. Hence

$$2 - 2\langle a, c\rangle = \|a-c\|^2 \leqslant \left(\|a-b\| + \|c-b\|\right)^2 \leqslant 8\delta\,.$$

■

Lemma 34 *Let $\mathcal{P} \subset \mathbb{R}^{d\times d}$ be some set of PSD matrices. For matrix $M \in \mathbb{R}^{n\times d}$ let*

$$\mathfrak{s}_{\mathcal{P}}(M) := \sup_{X\in\mathcal{P}} \left|\left\langle M^\top M, X\right\rangle\right|.$$

Then for arbitrary matrices $A, B \in \mathbb{R}^{n\times d}$,

$$\sup_{X\in\mathcal{P}} \left|\left\langle B^\top A + A^\top B, X\right\rangle\right| \leqslant 2\sqrt{\mathfrak{s}_{\mathcal{P}}(A)\cdot\mathfrak{s}_{\mathcal{P}}(B)}\,.$$

Proof Let X' be an arbitrary element of S. For some $c \in 0$ (we will choose the value of c later), let $C = (A - cB)^\top (A - cB)$. Since C and X' are PSD, $\langle C, X'\rangle \geqslant 0$. Hence

$$\left|c\cdot\left\langle B^\top A + A^\top B, X'\right\rangle\right| \leqslant \left\langle A^\top A, X'\right\rangle + c^2\left\langle B^\top B, X'\right\rangle \leqslant \mathfrak{s}_{\mathcal{P}}(A) + c^2\mathfrak{s}_{\mathcal{P}}(B)\,.$$

Therefore,

$$\left|\left\langle B^\top A + A^\top B, X'\right\rangle\right| \leqslant \frac{\mathfrak{s}_{\mathcal{P}}(A)}{|c|} + |c|\cdot\mathfrak{s}_{\mathcal{P}}(B)\,.$$

To minimize this expression, we can take $|c| = \sqrt{\frac{\mathfrak{s}_{\mathcal{P}}(A)}{\mathfrak{s}_{\mathcal{P}}(B)}}$. Hence

$$\left|\left\langle B^\top A + A^\top B, X'\right\rangle\right| \leqslant 2\sqrt{\mathfrak{s}_{\mathcal{P}}(A)\cdot\mathfrak{s}_{\mathcal{P}}(B)}\,.$$

Since it holds for arbitrary $X' \in \mathcal{S}$, we get the desired bound. ■

Appendix G. Concentration Inequalities

Fact 35 (Chernoff's inequality, Vershynin (2018)) *Let $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$ be independent Bernoulli random variables such that $\mathbb{P}(\boldsymbol{\zeta}_i = 1) = \mathbb{P}(\boldsymbol{\zeta}_i = 0) = p$. Then for every $\Delta > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n \boldsymbol{\zeta}_i \geqslant pn(1+\Delta)\right) \leqslant \left(\frac{e^{-\Delta}}{(1+\Delta)^{1+\Delta}}\right)^{pn}.$$

and for every $\Delta \in (0,1)$,

$$\mathbb{P}\left(\sum_{i=1}^n \boldsymbol{\zeta}_i \leqslant pn(1-\Delta)\right) \leqslant \left(\frac{e^{-\Delta}}{(1-\Delta)^{1-\Delta}}\right)^{pn}.$$

Fact 36 (Hoeffding's inequality, Wainwright (2019)) *Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be mutually independent random variables such that for each $i \in [n]$, $\boldsymbol{z}_i$ is supported on $[-c_i, c_i]$ for some $c_i \geqslant 0$. Then for all $t \geqslant 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}\,\boldsymbol{z}_i)\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right).$$

Fact 37 (Bernstein's inequality Wainwright (2019)) *Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ be mutually independent random variables such that for each $i \in [n]$, $\boldsymbol{z}_i$ is supported on $[-B, B]$ for some $B \geqslant 0$. Then for all $t \geqslant 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}\,\boldsymbol{z}_i) \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2\sum_{i=1}^n \mathbb{E}\,\boldsymbol{z}_i^2 + \frac{2Bt}{3}}\right).$$

Fact 38 *Wainwright (2019) Let $X \sim N(0, \sigma^2)$, then for all $t > 0$,*

$$\mathbb{P}(X \geqslant \sigma \cdot t) \leqslant e^{-t^2/2}\,.$$

Fact 39 *Laurent and Massart (2000)Let $X \sim \chi^2_m$, then for all $x > 0$,*

$$\mathbb{P}\big(X - m \geqslant 2x + 2\sqrt{mx}\big) \leqslant e^{-x}$$
$$\mathbb{P}(m - X \geqslant x) \leqslant e^{-\frac{x^2}{4m}}$$

Fact 40 *Wainwright (2019) Let $W \sim N(0,1)^{n\times d}$. Then with probability $1 - \exp(-t/2)$,*

$$\|W\| \leqslant \sqrt{n} + \sqrt{d} + \sqrt{t}$$

and

$$\left\|W^{\mathsf{T}}W - n\mathrm{Id}\right\| \leqslant d + 2\sqrt{dn} + t + 4\sqrt{t(n+d)}\,.$$

Fact 41 *d'Orsi et al. (2020) Let $\boldsymbol{W} \sim N(0,1)^{n\times d}$ be a Gaussian matrix. Let $1 \leqslant k \leqslant d$. Then with probability at least $1 - \left(\frac{k}{ed}\right)^k$*

$$\max_{\substack{u\in\mathbb{R}^n\\ \|u\|=1}} \max_{\substack{k\text{-sparse } v\in\mathbb{R}^d\\ \|v\|=1}} u^{\mathsf{T}}\boldsymbol{W}v \leqslant \sqrt{n} + 3\sqrt{k\ln\left(\frac{ed}{k}\right)}$$

and

$$\max_{\substack{k\text{-sparse } v\in\mathbb{R}^d\\ \|v\|=1}} v^\top \boldsymbol{W}^\top\boldsymbol{W}v - n \leqslant 10\sqrt{kn\ln(ed/k)} + 10k\ln(ed/k)\,.$$

Lemma 42 *Let $\boldsymbol{Y}$ be an instance of sparse PCA in Wishart model. For $m \in \mathbb{N}$ let $Z_m = \left\{z \in \{0,1\}^d \;\middle|\; \|z\|_1 \leqslant m\right\}$. Suppose that $m \geqslant 100\ln d$ and $n \geqslant 0.1\cdot m\ln(ed/m)$. Then, with probability at least $1 - d^{-10}$,*

$$\max_{z\in Z_m}\left\|\left(\boldsymbol{Y}^\top\boldsymbol{Y} - n\mathrm{Id} - \beta\|\boldsymbol{u}\|^2 vv^\top\right)\circ\left(zz^\top\right)\right\| \leqslant 10\sqrt{(n+\beta n)\cdot m\ln(ed/m)} + 10m\ln(ed/m)\,.$$

Proof We can write $\left(\boldsymbol{Y}^\top\boldsymbol{Y} - n\mathrm{Id} - \beta\|\boldsymbol{u}\|^2 vv^\top\right)\circ\left(zz^\top\right)$ as

$$\left(\boldsymbol{W}^\top\boldsymbol{W} - n\cdot\mathrm{Id} + \sqrt{\beta}\boldsymbol{W}^\top\boldsymbol{u}v^\top + \sqrt{\beta}v\boldsymbol{u}^\top\boldsymbol{W}\right)\circ\left(zz^\top\right).$$

Note that

$$\left\|\left(\boldsymbol{W}^\top\boldsymbol{u}v^\top + v\boldsymbol{u}^\top\boldsymbol{W}\right)\circ\left(zz^\top\right)\right\| \leqslant 2\left\|\left(\frac{1}{\|\boldsymbol{u}\|}\boldsymbol{u}^\top\boldsymbol{W}\right)\circ z\right\|\cdot\|\boldsymbol{u}\|\,.$$

With probability at least $1 - \exp(n/10)$, $\|\boldsymbol{u}\| \leqslant \sqrt{2n}$. By lemma 43, with probability at leaast $1 - \exp(-m)$,

$$\left\|\left(\boldsymbol{W}^\top\boldsymbol{u}v^\top + v\boldsymbol{u}^\top\boldsymbol{W}\right)\circ\left(zz^\top\right)\right\| \leqslant 5\sqrt{\beta n}\cdot\sqrt{m\ln(em/k)}\,.$$

By fact 40, for every $m' \in \mathbb{N}$, $\boldsymbol{G} \sim N(0,1)^{n \times m'}$ satisfies

$$\left\| \boldsymbol{G}^\top \boldsymbol{G} - n \cdot \mathrm{Id} \right\| \leqslant m' + 2\sqrt{m'n} + \tau + 4\sqrt{\tau(m'+n)}\,.$$

with probability $1 - \exp(-\tau/2)$ (for every $\tau > 0$). By union bound over all sets of size $m' \leqslant m$ we get the desired bound. ■

Lemma 43 *Let $\boldsymbol{W} \sim N(0,1)^{n \times d}$ and let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^d$ be vectors independent of $\boldsymbol{W}$. For $m \in \mathbb{N}$ let $Z_m = \left\{ z \in \{0,1\}^d \;\middle|\; \|z\|_1 \leqslant m \right\}$. Suppose that $m \geqslant 100 \ln d$. Then with probability at least $1 - d^{-10}$,*

$$\max_{z \in Z_m} \left\| \left(b a^\top \boldsymbol{W} \right) \circ \left(z z^\top \right) \right\| \leqslant 3 \cdot \|a\| \cdot \|b\| \cdot \sqrt{m \ln(em/k)}\,.$$

Proof For fixed set $J \subset [d]$, let $\mathbf{1}_J \in \{0,1\}^d$ be an indicator vector of this set. Random vector $\left(\frac{1}{\|\boldsymbol{u}\|} \boldsymbol{u}^\top \boldsymbol{W} \right) \circ \mathbf{1}_J$ has standard $|J|$-dimensional Gaussian distribution, hence by fact 40, for all $\tau > 0$, with probability at least $1 - \exp(-\tau/2)$,

$$\left\| \left(\frac{1}{\|a\|} a^\top \boldsymbol{W} \right) \circ \mathbf{1}_J \right\| \leqslant \sqrt{|J|} + 1 + \sqrt{\tau}\,.$$

By union bound over all sets of size of size at most m, we get

$$\max_{z \in Z_m} \left\| \left(\frac{1}{\|a\|} a^\top \boldsymbol{W} \right) \circ z \right\| \leqslant 3\sqrt{m \cdot \ln(ed/m)}$$

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