Sparse PCA

Sinho Chewi Patrik Robert Gerber

December 9, 2019

Contents

1 Introduction

Our goal is to give an overview of some theoretical guarantees for sparse PCA on a spiked covariance ensemble. We follow [Wai19] for the discussion of sparse PCA, and [Ver18] for matrix concentration inequalities.

First, we briefly explain what *principal components analysis (PCA)* is. Consider a random vector X in $\mathbb{R}^d$ which has mean zero. Its covariance matrix is $\Sigma := \mathbb{E}[XX^\mathsf{T}]$. Since Σ is symmetric (in fact, positive semidefinite), it has an eigendecomposition $\Sigma = \sum_{i=1}^d \sigma_i^2 u_i u_i^\mathsf{T}$, where $\sigma_1^2 \geq \cdots \geq \sigma_d^2 \geq 0$. PCA simply refers to computing these eigenvectors and using them for data analysis; in particular, one tends to truncate the eigendecomposition and keep only the eigenvectors corresponding to the largest eigenvalues. This is justified because the largest eigenvector has numerous interpretations (the other eigenvectors have similar interpretations):

- u_1 is the direction of largest variance, in the sense that

$$u_1 = \operatorname*{arg\,max}_{S^{d-1}} \operatorname{var} \langle X, \cdot \rangle;$$

- the outer product $u_1 u_1^\mathsf{T}$ is the best rank-one approximation to Σ, in the sense that $u_1 u_1^\mathsf{T} = \arg\min_{M \in \mathbb{R}^{d\times d},\, \operatorname{rank} M \le 1} |||M - \Sigma|||$, where $|||\cdot|||$ is any unitarily invariant matrix norm;
- among all one-dimensional subspaces S, the projection of X onto S minimizes the loss (expected squared error) when $S = \operatorname{span} u_1$.

Typically one does not have access to the true distribution of X, which is required to compute Σ (and thus u_1). Instead, one has access to i.i.d. samples $X_1, \dots, X_n$ drawn from the same distribution as X. In this case, we approximate Σ by the empirical covariance matrix $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\mathsf{T}$, and we approximate u_1 by computing $\hat{u}_1$, the principal eigenvector of $\hat{\Sigma}$.

In many problems, we are interested in the principal eigenvector because it encodes information about the structure of the data. For instance, one of the principal eigenvectors for the *community detection* problem encodes the structure of the communities; see [Ver18]. For this model and other similar models, PCA is also known as the *spectral method*, and we want to understand the statistical guarantees of the method; that is, how close is $\hat{u}_1$ to u_1?

Here, we will focus on a model referred to as the *spiked covariance ensemble*. In this model the i.i.d. datapoints $X_1, \dots, X_n$ are generated by

$$X_i = \sqrt{\nu}\xi_i u_1 + w_i, \tag{1.1}$$

where $\xi_i \in \mathbb{R}$ is a zero-mean unit-variance random variable and w_i is an independent zero-mean vector with covariance matrix I_d. We say that the spiked covariance ensemble has sub-Gaussian tails if both ξ_i and w_i are sub-Gaussian with variance proxy 1 for each $i \in [n]$. A simple calculation verifies that the covariance matrix of X then takes the form

$$\Sigma = I_d + \nu u_1 u_1^\mathsf{T}, \tag{1.2}$$

so that u_1 is the principal eigenvector. The quantity ν is the *eigengap* (the difference between the largest and second-largest eigenvalues) and it plays the role of a *signal-to-noise (SNR)* ratio for the problem.

2 Low-Dimensional PCA

In this section we discuss results for PCA in the low-dimensional setting i.e. the case when the ratio d/n is small (in a sense made precise later).

There are essentially two parts to the analysis that can be tackled separately.

- The first part is completely deterministic and involves matrix perturbation theory. Viewing $\hat{\Sigma}$ as a perturbation of Σ, how much can its eigenspaces differ from those of Σ?
- The second part of the analysis is to quantity exactly how much $\hat{\Sigma}$ deviates from Σ. This is the part of the analysis that address the stochasticity of the problem, and this is where we need tools from non-asymptotic random matrix theory.

2.1 A Deterministic Perturbation Argument

We will start with the deterministic perturbation argument. Since the principal eigenvector has the variational characterization $u_1 = \arg\max_{v \in S^{d-1}} \langle v, \Sigma v \rangle$, and likewise for $\hat{u}_1$ and $\hat{\Sigma}$, we can view $\hat{u}_1$ as the solution to an optimization problem. It is common in statistics to define estimators in this way, and the analysis of such estimators typically begins with the simple observation that

$$\langle \hat{u}_1, \hat{\Sigma} \hat{u}_1 \rangle \geq \langle u_1, \hat{\Sigma} u_1 \rangle.$$

This kind of inequality is known as a *basic inequality* in the statistics literature. Later, when we discuss sparse PCA, we will actually consider constrained estimators. To accomodate this general setting, let $\theta^* := \arg\max_{v \in \mathcal{C}} \langle v, \Sigma v \rangle$ and $\hat{\theta} := \arg\max_{v \in \mathcal{C}} \langle v, \hat{\Sigma} v \rangle$, where $\mathcal{C}$ is a general constraint set. The basic inequality now takes the form

$$\langle \hat{\theta}, \hat{\Sigma} \hat{\theta} \rangle \geq \langle \theta^*, \hat{\Sigma} \theta^* \rangle.$$

Manipulating this inequality yields the following lemma.

Lemma 2.1 ([Wai19, Lemma 8.1])**.** *Suppose Σ has eigengap $\nu > 0$ and define $\Psi(\Delta) := \langle \Delta, (\hat{\Sigma} - \Sigma)\Delta \rangle + 2\langle \Delta, (\hat{\Sigma} - \Sigma)\theta^* \rangle$. Then, it holds that*

$$\frac{\nu}{2} \|\hat{\theta} - \theta^*\|_2^2 \leq |\Psi(\hat{\theta} - \theta^*)|.$$

The point is that the RHS is easier to control because it involves bounding the perturbation $\hat{\Sigma} - \Sigma$ via non-asymptotic random matrix theory.

Let us write $\mathcal{S}^{d \times d}$ for the set of d-dimensional real symmetric matrices and $\mathcal{S}_+^{d \times d} \subset \mathcal{S}^{d \times d}$ for the subset of positive semidefinite matrices. Further, write $|||\cdot|||$ for the operator/largest eigenvalue norm on matrices. The following result is the workhorse of the analysis of PCA for the low-dimensional setting and it can be proved using Lemma 2.1.

Theorem 2.2 ([Wai19, Theorem 8.5])**.** *Let $\Sigma \in \mathcal{S}_+^{d \times d}$ have eigengap $\eta > 0$ and principal eigenvector θ^*. Suppose that P is a symmetric with $|||P||| < \nu/2$. Then the perturbed matrix $\hat{\Sigma} = \Sigma + P$ has a unique principal eigenvector $\hat{\theta}$ that satisfies the bound*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{2\|\tilde{p}\|_2}{\nu - 2|||P|||} \tag{2.3}$$

where the vector $\tilde{p}$ is defined by the relation

$$\tilde{P} := U^\mathsf{T} P U := \begin{pmatrix} \tilde{p}_{1,1} & \tilde{p}^\mathsf{T} \\ \tilde{p} & \tilde{P}_{2,2} \end{pmatrix} \tag{2.4}$$

and U is the orthonormal matrix with eigenvectors of Σ as its columns in decreasing order.

2.2 Application to the Spiked Covariance Ensemble

Let us now apply Theorem 2.2 to the specific case of the spiked covariance model introduced in (1.1). To be precise, let $X_1, \dots, X_n$ be an i.i.d. sample from a sub-Gaussian spiked covariance model that has covariance matrix Σ with eigengap $\nu > 0$ and principal eigenvector θ^*. Finally let $\hat{\Sigma}$ denote the sample covariance matrix. We have the following result:

Corollary 2.5 ([Wai19, Corollary 8.7]). *Suppose that* $\sqrt{\frac{d}{n}} \leq \frac{1}{128}\sqrt{\frac{\nu^2}{\nu+1}} \wedge 1$. *Then, with probability at least* $1 - C_1 \exp(-C_2 n \min\{\sqrt{\nu}\delta, \nu\delta^2\})$ *the matrix* $\hat{\Sigma}$ *has a unique principal eigenvector* $\hat{\theta}$ *such that*

$$\|\hat{\theta} - \theta^*\|_2 \leq C_0 \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{d}{n}} + \delta \tag{2.6}$$

where $C_0, C_1, C_2 > 0$ *are constants independent of* d, n *and* δ.

Remark 2.7. *Notice that the above statement holds only if* $d < n$.

3 High-Dimensional PCA

We saw that the result (Corollary 2.5) of the previous section held only when the number of dimensions was lower than the number of samples, in other words when $d < n$. It is a natural question to ask then: is it possible to bound the error between the true principal eigenvector θ^* and the sample principal eigenvector $\hat{\theta}$ even when $d > n$? As Wainwright puts it, the answer turns out to be a dramatic 'no'. In fact, if d/n is bounded away from zero, the vector $\hat{\theta}$ becomes asymptotically *orthogonal* to the true direction θ^*. This means that performing classical PCA in this regime is no better than projecting our data onto a random direction.

3.1 Sparse PCA

Not all is lost however. In many practical applications it is reasonable to assume that the principal eigenvector θ^* is *sparse*. This assumption encodes the idea that even for extremely high-dimensional data the 'intrinsic dimension' should be small.

Suppose that we know a priori that the principal eigenvector θ^* of Σ is s-sparse, meaning that

$$\|\theta^*\|_0 := \#\{i \in [d] \,:\, \theta_i^* \neq 0\} \leq s, \tag{3.1}$$

where we assume that $s \ll d$. Then it is natural to formulate sparse PCA as the following optimization problem:

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1}{\arg\max} \langle \theta, \hat{\Sigma}\theta \rangle \qquad \text{such that } \|\theta\|_0 \leq s. \tag{3.2}$$

The above optimization problem is combinatorial in nature and is unfortunately NP-hard. A formulation more amenable to computation is the one where the 0-norm is replaced by the 1-norm which is common practice in high-dimensional statistics. The reason for this is that such 1-norm regularization favors sparsity thereby replacing the hard, combinatorial constraint by a soft one. In what follows we will analyze the following penalized optimization problem:

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1}{\arg\max} \{\langle \theta, \hat{\Sigma}\theta \rangle - \lambda_n \|\theta\|_1\} \qquad \text{such that } \|\theta\|_1 \leq \nu \sqrt{\frac{n}{\log d}}, \tag{3.3}$$

where $\lambda_n \geq 0$ is a hyperparameter to be chosen later. Observe from the bound $\|\theta^*\|_1 \leq \sqrt{\|\theta^*\|_0}\|\theta^*\|_2$ that θ^* will be feasible for the program (3.3) provided that $\sqrt{s} \leq \nu \sqrt{\frac{n}{\log d}}$. We will henceforth assume that this is the case.

Remark 3.4. *It is also difficult to solve the program (3.3) because it is non-convex. There are SDP relaxations to this program (see [Wai19, Exercises 8.8-8.9]), and we will later discuss a diffrent implementable algorithm for the sparse PCA problem. For now, our interest in the program (3.3) stems from a desire to understand, information-theoretically, how well a simple regularized estimator can perform.*

Suppose again that we have an i.i.d. sample $X_1, \ldots, X_n$ from a sub-Gaussian spiked covariance ensemble with covariance matrix Σ, s-sparse principal eigenvector θ^* and eigengap $\nu > 0$. Our goal will be to prove the following:

Theorem 3.5 ([Wai19, Corollary 8.12])**.** *There exists universal (i.e., not depending on n or d) constants $c, c_1, c_2, c_3, c_4 > 0$ such that for any n with $\frac{s \log d}{n} \leq c \min\{1, \frac{\nu^2}{1+\nu}\}$ and any $\delta \in (0,1)$, every solution $\hat{\theta}$ to the problem (3.3) with $\lambda_n = c_1 \sqrt{\nu + 1}\{\sqrt{\frac{\log d}{n}} + \delta\}$ satisfies*

$$\|\hat{\theta} - \theta^*\|_2 \wedge \|\hat{\theta} + \theta^*\|_2 \leq c_2 \sqrt{\frac{\nu+1}{\nu^2}} \Big\{ \sqrt{\frac{s \log d}{n}} + \delta \Big\} \tag{3.6}$$

with probability at least $1 - c_3 \exp\big(-c_4 \frac{n}{s} \min\{\nu, \nu^2, \delta^2\}\big)$.

The strategy to prove Theorem 3.5 is similar to that outlines in Section 2.1. Accordingly the next section concerns itself with a deterministic result about the solutions to (3.3).

3.2 A Second Deterministic Argument

Once again, the first step is to write down the basic inequality associated with (3.3), which is

$$\langle \hat{\theta}, \hat{\Sigma}\hat{\theta} \rangle - \lambda_n \|\hat{\theta}\|_1 \geq \langle \theta^*, \hat{\Sigma}\theta^* \rangle - \lambda_n \|\theta^*\|_1.$$

If we perform the same manipulations that are used to derive Lemma 2.1, then we obtain the inequality

$$\frac{\nu}{2}\|\hat{\theta}-\theta^*\|_2^2 - |\Psi(\hat{\theta}-\theta^*)| \le \lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}. \tag{3.7}$$

Observe that the only difference between this inequality and Lemma 2.1 is the presence of the term $\lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}$ on the RHS, which comes from the regularization term.

From now on, denote $\hat{\Delta} := \hat{\theta} - \theta^*$ the residual vector. Let $S \subseteq [d]$ denote the support of θ^*, let $S^c := [d] \setminus S$, and for a vector $v \in \mathbb{R}^d$, let v_S denote the restriction of v to the coordinates in S, that is, $(v_S)_i = v_i$ if $i \in S$ and $(v_S)_i = 0$ otherwise. We first give a simple upper bound the RHS of (3.7):

$$\lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\} = \lambda_n\{\|\theta^*_S\|_1 - \|\hat{\theta}_S\|_1 - \|\hat{\theta}_{S^c}\|_1\} \le \lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\}. \tag{3.8}$$

Again, the crux of the argument decomposes into two main steps:

- We will argue that the residual $\hat{\Delta}$ satisfies the inequality $\|\hat{\Delta}\|_1 \le 4\sqrt{s}\|\hat{\Delta}\|_2$. The proof of this step relies on the regularization that we added to the program (3.3).

  To understand intuitively what this inequality means, observe that for any vector $v \in \mathbb{R}^d$ we have the inequality $\|v\|_1 \le \sqrt{d}\|v\|_2$, where equality is attained by a non-sparse vector (e.g., the vector $(1, \dots, 1)$). If v is s-sparse, then we have the better inequality $\|v\|_1 \le \sqrt{s}\|v\|_2$, and in the extreme case where $v = e_1$, say, then $\|v\|_1 = \|v\|_2$. Thus, the inequality $\|\hat{\Delta}\|_1 \le 4\sqrt{s}\|\hat{\Delta}\|_2$ says that our residual is *analytically sparse* (that is, it may not be exactly sparse but it shares many of the analytic properties of a sparse vector).

- Next, we will use tools from random matrix theory to prove an upper bound on $|\Psi(\hat{\Delta})|$. Combining this bound with the inequality $\|\hat{\Delta}\|_1 \le 4\sqrt{s}\|\hat{\Delta}\|_2$ from the first step, along with the basic inequality (3.7), we will derive an upper bound on the error $\|\hat{\Delta}\|_2$.

We end this section by completing the first step:

Lemma 3.9. *Suppose that the regularization level λ_n is chosen so that*

$$\lambda_n\|\hat{\Delta}\|_1 \ge 2\Big\{|\Psi(\hat{\Delta})| - \frac{\nu}{2}\|\hat{\Delta}\|_2^2\Big\}. \tag{3.10}$$

Then, it holds that $\|\hat{\Delta}_{S^c}\|_1 \le 3\|\hat{\Delta}_S\|_1$. In particular, it implies the inequality $\|\hat{\Delta}\|_1 \le 4\sqrt{s}\|\hat{\Delta}\|_2$.

Proof. From the basic inequality (3.7) and (3.8), as well as the assumption (3.10) on the regularization, we have the bound

$$0 \le |\Psi(\hat{\Delta})| - \frac{\nu}{2}\|\hat{\Delta}\|_2^2 + \lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\}$$

$$\leq \frac{\lambda_n}{2}\|\hat{\Delta}\|_1 + \lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\} = \lambda_n\Big\{\frac{3}{2}\|\hat{\Delta}_S\|_1 - \frac{1}{2}\|\hat{\Delta}_{S^c}\|_1\Big\}.$$

This proves the first inequality. To prove the second inequality,

$$\|\hat{\Delta}\|_1 = \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \leq 4\|\hat{\Delta}_S\|_1 \leq 4\sqrt{s}\|\hat{\Delta}\|_2. \qquad \square$$

The condition (3.10) dictates our choice of the regularization parameter λ_n. In the next section, we will obtain an upper bound on $|\Psi(\hat{\Delta})|$, which will show that choosing λ_n as in Theorem 3.5, condition (3.10) will be met w.h.p.

3.3 Controlling the Randomness

From now on, we will use the convention of using c and C to denote unspecified universal positive constants which are allowed to change from line to line.[1]

Let $P := \hat{\Sigma} - \Sigma$ be the perturbation of the covariance matrix and recall that by definition, $\Psi(\Delta; P) := \langle \Delta, P\Delta\rangle + 2\langle \Delta, P\theta^*\rangle$; here, we explicitly incorporate the dependence of Ψ on P into the notation. In the spiked model (1.1), let $\bar{w} := \frac{1}{n}\sum_{i=1}^n \xi_i w_i$. The covariance matrix and sample covariance matrix are

$$\Sigma = I_d + \nu\theta^*(\theta^*)^\mathsf{T},$$
$$\hat{\Sigma} = \frac{\nu}{n}\sum_{i=1}^n \xi_i^2\theta^*(\theta^*)^\mathsf{T} + \sqrt{\nu}\{\bar{w}(\theta^*)^\mathsf{T} + \theta^*\bar{w}^\mathsf{T}\} + \frac{1}{n}\sum_{i=1}^n w_i w_i^\mathsf{T}.$$

Therefore, the perturbation splits into the three terms

$$P = \underbrace{\nu\Big(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - 1\Big)\theta^*(\theta^*)^\mathsf{T}}_{P_1} + \underbrace{\sqrt{\nu}\{\bar{w}(\theta^*)^\mathsf{T} + \theta^*\bar{w}^\mathsf{T}\}}_{P_2} + \underbrace{\frac{1}{n}\sum_{i=1}^n w_i w_i^\mathsf{T} - I_d}_{P_3}.$$

Accordingly, we split up $\Psi(\hat{\Delta}; P) = \Psi(\hat{\Delta}; P_1) + \Psi(\hat{\Delta}; P_2) + \Psi(\hat{\Delta}; P_3)$ and control each term separately. The main tool is concentration, and we will only sketch some of the arguments using scalar concentration bounds. Instead, we will focus our attention on the third term, which will require matrix concentration. Below, it will be useful to note that $|\langle\hat{\Delta}, \theta^*\rangle| = |1 - \langle\hat{\theta}, \theta^*\rangle| = \frac{1}{2}\|\hat{\theta} - \theta^*\|_2^2 = \frac{1}{2}\|\hat{\Delta}\|_2^2$.

For the first term, with probability at least $1 - 2\exp(-cn)$, it holds that $\big|\frac{1}{n}\sum_{i=1}^n \xi_i^2 - 1\big| \leq \frac{1}{48}$.[2] Therefore,

$$|\Psi(\hat{\Delta}; P_1)| \leq \nu\Big|\frac{1}{n}\sum_{i=1}^n \xi_i^2 - 1\Big|(\langle\hat{\Delta}, \theta^*\rangle^2 + 2|\langle\hat{\Delta}, \theta^*\rangle|) \leq \frac{\nu}{16}|\langle\hat{\Delta}, \theta^*\rangle| = \frac{\nu}{32}\|\hat{\Delta}\|_2^2.$$

For the second term,

$$|\Psi(\hat{\Delta}; P_2)| \leq 2\sqrt{\nu}|\langle\hat{\Delta}, \bar{w}\rangle\langle\hat{\Delta}, \theta^*\rangle + \langle\Delta, \bar{w}\rangle + \langle\theta^*, \bar{w}\rangle\langle\hat{\Delta}, \theta^*\rangle|$$

[1] Generally, we will try to use c to mean "a constant sufficiently small" and C for "a constant sufficiently large".

[2] See the discussion of concentration of the norm in [Ver18].

$$\leq 4\sqrt{\nu}\|\bar{w}\|_\infty\|\hat{\Delta}\|_1 + \sqrt{\nu}|\langle\theta^*, \bar{w}\rangle|\|\hat{\Delta}\|_2^2.$$

These terms are also controlled by concentration bounds. For $\delta > 0$ small, with high probability, $\|\bar{w}\|_\infty \leq C\sqrt{\frac{\log d}{n}} + \delta$ and $|\langle\theta^*, \bar{w}\rangle| \leq \frac{\sqrt{\nu}}{32}$.[3] Plugging these bounds in yields control of the second term:

$$|\Psi(\hat{\Delta}; P_2)| \leq \frac{\nu}{32}\|\hat{\Delta}\|_2^2 + C\sqrt{\nu}\Big\{\sqrt{\frac{\log d}{n}} + \delta\Big\}\|\hat{\Delta}\|_1.$$

For the third term,

$$|\Psi(\hat{\Delta}; P_3)| \leq |\langle\hat{\Delta}, P_3\hat{\Delta}\rangle| + 2|\langle\hat{\Delta}, P_3\theta^*\rangle| \leq |\langle\hat{\Delta}, P_3\hat{\Delta}\rangle| + 2\|P_3\theta^*\|_\infty\|\hat{\Delta}\|_1.$$

Similarly as before, we can control $\|P_3\theta^*\|_\infty \leq C\sqrt{\frac{\log d}{n}} + \delta$. It remains to bound the term $|\langle\hat{\Delta}, P_3\hat{\Delta}\rangle|$. Since this term requires the most care (as well as requiring the most tools from random matrix theory), we focus most of our attention on controlling this term.

The goal will be to prove:

Lemma 3.11. *With probability at least $1 - C\exp(-cn\min\{\nu, \nu^2\})$, it holds that for all $\Delta \in \mathbb{R}^d$,*

$$|\langle\Delta, P_3\Delta\rangle| \leq \frac{\nu}{16}\|\Delta\|_2^2 + \frac{C}{\nu}\frac{\log d}{n}\|\Delta\|_1^2.$$

Note that the bound is uniform over $\Delta \in \mathbb{R}^d$.

We will prove Lemma 3.11 in the next section, but first let us collect together the pieces to prove Theorem 3.5. From the bounds we have derived (including Lemma 3.11), we have

$$|\Psi(\hat{\Delta}; P)| \leq \frac{\nu}{8}\|\hat{\Delta}\|_2^2 + C\sqrt{\nu}\Big\{\sqrt{\frac{\log d}{n}} + \delta\Big\}\|\hat{\Delta}\|_1 + \frac{C}{\nu}\frac{\log d}{n}\|\hat{\Delta}\|_1^2 \tag{3.12}$$

$$\leq \frac{\nu}{8}\|\hat{\Delta}\|_2^2 + C\sqrt{\nu+1}\Big\{\sqrt{\frac{\log d}{n}} + \delta\Big\}\|\hat{\Delta}\|_1, \tag{3.13}$$

where we use the fact that $\hat{\theta}$, θ^* are feasible for the program (3.3), and thus $\|\hat{\Delta}\|_1 \leq 2\nu\sqrt{\frac{n}{\log d}}$.

Proof of Theorem 3.5. Recall that in order to apply Lemma 3.9, we must choose the regularization parameter λ_n to satisfy condition (3.10). The bound (3.13)

[3] We sketch some details here; see [Wai19; Ver18]. The random variable $(\xi_1 w_1)_1$ is subexponential with parameter $O(1)$, so $\frac{1}{n}\sum_{i=1}^n (\xi_i w_i)_1$ is subexponential with parameter $O(\frac{1}{\sqrt{n}})$. It implies that for $u \geq 0$ small, $\mathbb{P}\{\frac{1}{n}\sum_{i=1}^n (\xi_i w_i)_1 \geq u\} \leq \exp(-cnu^2)$. Taking a union bound and setting $u = C\sqrt{\frac{\log d}{n}} + t$ yields $\mathbb{P}\{\|\bar{w}\|_\infty \geq C\sqrt{\frac{\log d}{n}} + t\} \leq \exp(-cn\delta^2)$.

Similarly, $|\langle\theta^*, \bar{w}\rangle|$ is subexponential with parameter $O(\frac{1}{\sqrt{n}})$, so for small $u \geq 0$, it holds that $\mathbb{P}\{|\langle\theta^*, \bar{w}\rangle| \geq \frac{\sqrt{\nu}}{32}\} \leq 2\exp(-cn\nu)$.

tells us that we may choose the regularization

$$\lambda_n = C\sqrt{\nu+1}\Big\{\sqrt{\frac{\log d}{n}} + \delta\Big\} \tag{3.14}$$

and the condition (3.10) will be met.

Now, we have the bound

$$\begin{aligned}
\frac{\nu}{2}\|\hat{\Delta}\|_2^2 &\le |\Psi(\hat{\Delta})| + \lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\} && \text{by (3.7), (3.8)} \\
&\le |\Psi(\hat{\Delta})| + \lambda_n\|\hat{\Delta}\|_1 \\
&\le \frac{\nu}{8}\|\hat{\Delta}\|_2^2 + C\sqrt{\nu+1}\Big\{\sqrt{\frac{\log d}{n}} + \delta\Big\}\|\hat{\Delta}\|_1 + \lambda_n\|\hat{\Delta}\|_1 && \text{by (3.13)} \\
&\le \frac{\nu}{8}\|\hat{\Delta}\|_2^2 + C\sqrt{\nu+1}\Big\{\sqrt{\frac{\log d}{n}} + \delta\Big\}\|\hat{\Delta}\|_1 && \text{by (3.14)} \\
&\le \frac{\nu}{8}\|\hat{\Delta}\|_2^2 + C\sqrt{s}\sqrt{\nu+1}\Big\{\sqrt{\frac{\log d}{n}} + \delta\Big\}\|\hat{\Delta}\|_2 && \text{by Lemma 3.10}
\end{aligned}$$

which holds with probability at least $1 - C\exp(-cn\min\{\nu, \nu^2, \delta^2\})$. Rearranging this bound and replacing δ with $\delta/\sqrt{s}$ yields the conclusion

$$\|\hat{\Delta}\|_2 \le C\sqrt{\frac{\nu+1}{\nu^2}}\Big\{\sqrt{\frac{s\log d}{n}} + \delta\Big\}$$

with probability at least $1 - C\exp\big(-c\frac{n}{s}\min\{\nu, \nu^2, \delta^2\}\big)$. □

It only remains to prove the matrix concentration bound in Lemma 3.11.

3.4 Restricted Isometry

In this section, our goal is to prove Lemma 3.11, which bounds $|\langle\Delta, P_3\Delta\rangle|$ for all $\Delta \in \mathbb{R}^d$. First, let us look at a naïve way of bounding this quantity. Recall that $P_3 = \frac{1}{n}\sum_{i=1}^n w_i w_i^\mathsf{T} - I_d$, so the result on estimation of a covariance matrix in [Ver18, Theorem 4.7.1] tells us that, at best, we can expect $|||P||| \le C\frac{d}{n}$. This yields the bound $|\langle\Delta, P_3\Delta\rangle| \le C\frac{d}{n}\|\Delta\|_2^2$, which is far weaker than Lemma 3.11.[4] In fact, the $\frac{d}{n}$ scaling is exactly what showed up in Theorem 2.2, and the whole point of considering sparse PCA is to avoid this so-called "curse of dimensionality"! We will need to better leverage the sparsity.

Specifically, if we look at Lemma 3.9, then we see that $\hat{\Delta}$ will eventually be analytically sparse, in the sense that $\|\hat{\Delta}\|_1 \le 4\sqrt{s}\|\hat{\Delta}\|_2$, and our hope is to replace the $\frac{d}{n}$ scaling with $\sqrt{\frac{s\log d}{n}}$. *Note, however*, that Lemma 3.11 is used to *prove* Lemma 3.9, so we cannot make the *a priori* assumption that Δ is

[4] For a non-sparse vector Δ, the bound in Lemma 3.11 is of order $\frac{d\log d}{n}\|\Delta\|_2^2$. When we say that Lemma 3.11 is a stronger bound, we mean that it is much tighter than $\frac{d}{n}\|\Delta\|_2^2$ for *sparse* (or nearly sparse) vectors.

analytically sparse. We must prove a bound on the quantity $|\langle \Delta, P_3 \Delta \rangle|$ for an *arbitrary* $\Delta \in \mathbb{R}^d$. Nevertheless, this discussion tells us that it is not a bad idea to first look at the case when Δ is sparse.

In fact, suppose that the first s coordinates of Δ are non-zero, and the remaining coordinates are zero. Let $S = [s]$ denote the first s coordinates. Then, $\langle \Delta, P_3 \Delta \rangle = \langle \Delta_S, (P_3)_S \Delta_S \rangle$, where $(P_3)_S$ denotes the restriction of P_3 to the rows and columns corresponding to the index set S. We have effectively reduced the dimensionality of P_3, so the $C\sqrt{\frac{d}{n}}$ bound for the operator norm of P_3 can be replaced by the operator norm bound $C\sqrt{\frac{s}{n}}$ for $(P_3)_S$. To be precise, if we invoke [Ver18, Exercise 4.7.3] or [Wai19, Theorem 6.2], then we obtain that for small $\alpha > 0$, with probability at least $1 - C\exp(-cn\alpha^2\nu^2)$, it holds that $|||(P_3)_S||| \le C\sqrt{\frac{s}{n}} + \alpha\nu$.

Now, we take a union bound over the $\binom{d}{s}$ subsets of s coordinates. Recall that the sparsity of θ^* is at most $\nu^2 \frac{n}{\log d}$, so we consider $s \approx c\nu^2 \frac{n}{\log d}$. From this, we have $\log \binom{d}{s} \le s \log d \approx c\nu^2 n$. So, provided we choose c to be a small multiple of α^2, we will have with probability at least $1 - C\exp(-cn\alpha^2\nu^2)$ that $\max_{S \in \binom{[d]}{s}} |||(P_3)_S||| \le C\sqrt{\frac{s}{n}} + \alpha\nu \le C\alpha\nu$.

Remark 3.15. *Observant readers will have noticed that the above bound implies that the matrix* $\frac{1}{n}\sum_{i=1}^n w_i w_i^\mathsf{T}$ *acts as an approximate isometry on* s*-sparse vectors. This property is called the* restricted isometry property (RIP)*, and plays a prominent role in sparse signal recovery [BG11; Kol11; Moi18; Ver18; Wai19]. It is not too surprising that it appears in this problem, where we are trying to recover a sparse eigenvector.*

Returning to the proof, our bound immediately implies that for an s-sparse vector Δ (with $s \approx c\alpha^2\nu^2 \frac{n}{\log d}$), it holds that $|\langle \Delta, P_3 \Delta \rangle| \le C\alpha\nu\|\Delta\|_2^2$. This is quite a good bound, but we will need to extend it for vectors which are not exactly s-sparse. Suppose instead that $\|\Delta\|_1 \le C\sqrt{s}\|\Delta\|_2$; we will now use a standard technique for extending the bound to this case.

Let I_1 denote the indices corresponding to the s largest coordinates of Δ; let I_2 denote the indices corresponding to the next s largest coordinates of Δ; continuing in this manner, define the index sets $I_3, \dots, I_k$, where $k = \lceil \frac{d}{s} \rceil$. We now decompose $P_3\Delta = \sum_{i=1}^k P_3 \Delta_{I_k}$, and we prove some properties of this decomposition.

1. From our bound on $|||(P_3)_{I_1}|||$: $\|P_3\Delta_{I_1}\|_2 \le C\alpha\nu\|\Delta_{I_1}\|_2 \le C\alpha\nu\|\Delta\|_2$.
2. For each $i = 2, \dots, k$, by construction, the magnitude of every entry of Δ_{I_i} is bounded by the average of the magnitudes of the entries of $\Delta_{I_{i-1}}$ (i.e., $\frac{1}{s}\|\Delta_{I_{i-1}}\|_1$). This leads to the bound $\|\Delta_{I_i}\|_2 \le \frac{1}{\sqrt{s}}\|\Delta_{I_{i-1}}\|_1$. As before, it implies $\|P_3\Delta_{I_i}\|_2 \le C\alpha\nu\frac{1}{\sqrt{s}}\|\Delta_{I_{i-1}}\|_1$.

Summing up these bounds yields

$$\|P_3\Delta\|_2 \le \sum_{i=1}^k \|P_3\Delta_{I_i}\|_2 \le C\alpha\nu\Big\{\|\Delta\|_2 + \frac{1}{\sqrt{s}}\sum_{i=2}^k \|\Delta_{I_{i-1}}\|_1\Big\}$$

$$\leq C\alpha\nu\Big\{\|\Delta\|_2 + \frac{1}{\sqrt{s}}\|\Delta\|_1\Big\} \leq C\alpha\nu\|\Delta\|_2,$$

where we used the analytical sparsity of Δ. This implies $|\langle\Delta, P_3\Delta\rangle| \leq C\alpha\nu\|\Delta\|_2^2$.

Finally, suppose that $\|\Delta\|_1 \geq c\sqrt{s}\|\Delta\|_2$. In this case, we use the same decomposition as before, but we leave everything in terms of the ℓ_1-norm instead:

$$\begin{aligned}\|P_3^{1/2}\Delta\|_2 &\leq \sum_{i=1}^{k}\|P_3^{1/2}\Delta_{I_i}\|_2 \leq C\sqrt{\alpha\nu}\Big\{\|\Delta\|_2 + \frac{1}{\sqrt{s}}\sum_{i=2}^{k}\|\Delta_{I_{i-1}}\|_1\Big\} \\ &\leq C\sqrt{\alpha\nu}\Big\{\|\Delta\|_2 + \frac{1}{\sqrt{s}}\|\Delta\|_1\Big\} \leq C\sqrt{\frac{\alpha\nu}{s}}\|\Delta\|_1.\end{aligned}$$

Squaring this inequality yields $|\langle\Delta, P_3\Delta\rangle| \leq C\alpha\nu\frac{1}{s}\|\Delta\|_1^2$.

If we combine all of these bounds together and recall our choice of s, then we have proven the following inequality that holds for all $\Delta \in \mathbb{R}^d$:

$$|\langle\Delta, P_3\Delta\rangle| \leq C\alpha\nu\Big\{\|\Delta\|_2^2 + \frac{1}{\alpha^2\nu^2}\frac{\log d}{n}\|\Delta\|_1^2\Big\}.$$

Now it remains to choose α to be a sufficiently small absolute constant to make the constant in front of $\|\Delta\|_2^2$ equal to $\frac{1}{16}$, and this concludes the proof of Lemma 3.11.

3.5 Implementation Details

A wealth of algorithms have been proposed to compute the sparse principal components in (3.2). Various variants on the idea of thresholding produce good results while recently methods involving semidefinite relaxation have gained popularity. The early paper [JL09] proposed a very simple thresholding algorithm where the support of the unknown sparse principal vector is estimated by taking the features with largest sample variance. This "diagonal thresholding" algorithm is consistent in the regime where $s \leq O\big(\sqrt{\frac{n}{\log d}}\big)$. The information theoretic limit was shown to be $s \geq \Omega\big(\frac{n}{\log d}\big)$ in [AW08] meaning that for $s \geq \Omega\big(\frac{n}{\log d}\big)$ no algorithm (efficient or not) exists that computes the sparse principal eigenvector consistently. It was conjectured that solutions to the semidefinite relaxations might close the gap between the two bounds above, but in [KNV+15] the authors show that this is not the case. They also suggest a slightly more sophisticated thresholding algorithm that empirically performs at least as well as the semidefinite approach. Considering the above, we chose to implement the diagonal thresholding algorithm as described in [JL09] due to its simplicity and good performance.

References

[AW08] Arash A Amini and Martin J Wainwright. "High-dimensional analysis of semidefinite relaxations for sparse principal components". In: *2008 IEEE International Symposium on Information Theory*. IEEE. 2008, pp. 2454–2458.

[BG11] Peter Bühlmann and Sara van de Geer. *Statistics for high dimensional data*. Springer Series in Statistics. Methods, theory and applications. Springer, Heidelberg, 2011, pp. xviii+556.

[JL09] Iain M Johnstone and Arthur Yu Lu. "On consistency and sparsity for principal components analysis in high dimensions". In: *Journal of the American Statistical Association* 104.486 (2009), pp. 682–693.

[KNV+15] Robert Krauthgamer, Boaz Nadler, Dan Vilenchik, et al. "Do semidefinite relaxations solve sparse PCA up to the information limit?" In: *The Annals of Statistics* 43.3 (2015), pp. 1300–1322.

[Kol11] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Vol. 2033. Lecture Notes in Mathematics. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. Springer, Heidelberg, 2011.

[Moi18] Ankur Moitra. *Algorithmic aspects of machine learning*. Cambridge University Press, 2018.

[Ver18] Roman Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018, pp. xiv+284.

[Wai19] Martin J. Wainwright. *High-dimensional statistics*. Vol. 48. Cambridge Series in Statistical and Probabilistic Mathematics. A non-asymptotic viewpoint. Cambridge University Press, Cambridge, 2019, pp. xvii+552.