Entrywise Recovery Guarantees for Sparse PCA via Sparsistent Algorithms

Joshua Agterberg
Johns Hopkins University

Jeremias Sulam
Johns Hopkins University

Abstract

Sparse Principal Component Analysis (PCA) is a prevalent tool across a plethora of subfields of applied statistics. While several results have characterized the recovery error of the principal eigenvectors, these are typically in spectral or Frobenius norms. In this paper, we provide entrywise $\ell_{2,\infty}$ bounds for Sparse PCA under a general high-dimensional subgaussian design. In particular, our results hold for any algorithm that selects the correct support with high probability, those that are *sparsistent*. Our bound improves upon known results by providing a finer characterization of the estimation error, and our proof uses techniques recently developed for entrywise subspace perturbation theory.

1 INTRODUCTION

Principal component analysis (PCA) is a standard statistical technique for dimensionality reduction of data in an unsupervised manner. Given i.i.d mean-zero observations $X_1, \ldots, X_n \in \mathbb{R}^p$ with covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, the goal of PCA is to estimate the leading k-dimensional subspace of Σ, which has the interpretation of representing each observation as a linear combination of *principal components*, where each principal component is a direction of maximal variance. The classical theory of PCA (e.g. Anderson (2003)) shows that if the number of covariates p is fixed and the number of samples n tends to infinity, then the leading eigenvectors of the sample covariance approximate the leading eigenvectors of the population covariance well.

In the modern era of big data, it is often unrealistic to assume that p remains fixed in n. In the seminal work of Johnstone and Lu (2009), the authors introduced the *spiked covariance model* where the leading eigenvalue of the population covariance satisfies $\lambda_1 > 1$, while all other eigenvalues are all 1. In Johnstone and Lu (2009), the authors showed that if $\hat{u}_1$ is the leading eigenvector of the sample covariance and u_1 is the leading eigenvector of the population covariance, then $\langle \hat{u}_1, u_1 \rangle$ need not tend to 1 as p and n tend to infinity unless either $p/n \to 0$ or the leading eigenvalue λ_1 tends to infinity. They then went on to show that if λ_1 remains bounded away from infinity but the leading eigenvector is *sparse* then a simple thresholding estimator could yield consistent estimation. Since then, there has been much work on generalizing the model in Johnstone and Lu (2009) to settings where either the leading eigenvalues tend to infinity (Bao et al., 2020; Cai et al., 2020, 2021; Fan et al., 2020; Yan et al., 2021) or the leading eigenvectors are sparse (Amini and Wainwright, 2009; d'Aspremont et al., 2007; Cai et al., 2013; Gao et al., 2017; Gataric et al., 2020; Gu et al., 2014; Lei and Vu, 2015; Ma, 2013; Yang et al., 2015).

In this paper we consider the setting where the leading eigenvalues of the covariance matrix are bounded away from zero and infinity, but the leading k eigenvectors are s-sparse as n and p tend to infinity. There have been substantial theoretical (Banks et al., 2018; Cai et al., 2013; Krauthgamer et al., 2015; Vu and Lei, 2013; Wang et al., 2016) and methodological (Berthet and Rigollet, 2013; Chen and Rohe, 2020; Gataric et al., 2020; Ma, 2013; Rohe and Zeng, 2020; Xie et al., 2019) developments in sparse PCA. In Vu et al. (2013) the authors propose a semidefinite program enforcing sparsity to estimate the leading eigenvectors of the population covariance given only the sample covariance, and in Lei and Vu (2015) the authors provide general results for which the algorithm in Vu et al. (2013) selects the correct support. Similarly, Gu et al. (2014) propose a nonconvex algorithm that selects the correct support with high probability.

In many of the existing theoretical results on sparse PCA, authors are primarily concerned with subspace

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

estimation error in spectral or Frobenius norm (e.g. Cai et al. (2013); Vu et al. (2013); Vu and Lei (2013)). However, in many situations entrywise guarantees can lead to more refined results which can be useful for downstream inference. In this paper, building upon a host of recent works on entrywise guarantees for eigenvectors (Abbe et al., 2020a,b; Agterberg et al., 2021; Cai et al., 2021; Cape et al., 2019a,b; Charisopoulos et al., 2020; Chen et al., 2020a; Damle and Sun, 2020; Fan et al., 2018; Jin et al., 2019; Lei, 2019; Mao et al., 2020; Xia and Yuan, 2020; Xie et al., 2019; Xie, 2021; Yan et al., 2021), we study entrywise guarantees for sparse PCA for a very general class of models. Our main results hold for any *sparsistent* algorithm, i.e. one that selects the correct support for the eigenvectors, with high probability. Sparsistency has also been studied in other contexts in high-dimensional statistics, such as in sparse linear models (Fan and Li, 2001; Wainwright, 2009; Zhao and Yu, 2006). See Bühlmann and van de Geer (2011) for a more comprehensive overview.

The literature on entrywise eigenvector analysis includes a suite of tools and techniques to bound the entries of eigenvectors in ways that classical matrix perturbation theory (e.g. Horn and Johnson (2012); Stewart and Sun (1990); Bhatia (1997)) fails to address. The Davis-Kahan Theorem (Yu et al., 2014) provides a useful benchmark for eigenvector analysis, but this can lead to suboptimal entrywise bounds. The primary reason for the lack of optimality is due to the fact that the Davis-Kahan Theorem can be somewhat coarse, as it fails to take into account the probabilistic nature of empirical eigenvectors in statistical settings. Therefore, entrywise eigenvector bounds require careful probabilistic and matrix analysis techniques that go beyond what the Davis-Kahan Theorem and classical matrix perturbation theory can do. See Chen et al. (2020a) for an accessible introduction to entrywise eigenvector estimation. The only other work on entrywise eigenvector analysis in sparse PCA is in Xie et al. (2019), which is a Bayesian setting under the relatively stringent spiked model. In this paper we develop entrywise bounds for sparse PCA under a much more general model class. More specifically, our results hold for models satisfying a mild eigengap requirement (see Assumption 4) that includes the spiked model.

The rest of this paper is organized as follows. In Section 2 we provide the requisite background for sparse PCA and existing results on sparsistency. In Section 3 we provide our main results, and Section 4 includes the discussion. We include a sketch of our main proof in Section A, but the full proofs are relegated to the supplementary material.

1.1 Notation

We use capital letters to denote matrices and random vectors, which will be clear from context, and lower case letters to denote fixed vectors. We let $X_1, \ldots, X_n$ denote a collection of n random variables in $\mathbb{R}^p$. For a generic real-valued random variable X, its ψ_α *Orlicz norm of order* α (or just ψ_α norm) is defined via $\|X\|_{\psi_\alpha} := \inf\{t > 0 : \mathbb{E}\exp(|X|^\alpha/t) \leq 1\}$. Random variables with finite ψ_2 norm are called *subgaussian* and random variables with finite ψ_1 norm are called *subexponential*. More discussion on Orlicz norms is included in Appendix C in the supplementary material.

For $d_1 \geq d_2$, we define the set of matrices $U \in \mathbb{R}^{d_1 \times d_2}$ with orthonormal columns as $\mathbb{O}(d_1, d_2)$ and when $d = d_1 = d_2$, we denote this set as $\mathbb{O}(d)$. We use $\|\cdot\|$ as the spectral norm on matrices and the Euclidean norm on vectors, $\|\cdot\|_F$ as the Frobenius norm, and $\|\cdot\|_{\max}$ for the maximum entry norm. Except for the spectral norm, we write $\|\cdot\|_{p\to q}$ as the operator norm from $\ell_p \to \ell_q$; that is $\|M\|_{p\to q} := \sup_{\|x\|_p=1} \|Mx\|_q$. Of particular importance is the $2 \to \infty$ norm, which is the maximum row norm of a matrix. Except for the maximum entry norm, we write $\|\cdot\|_p$ to denote the entrywise p norm of a matrix viewed as a long vector. For a matrix M, $\mathrm{diag}(M)$ extracts its diagonal, and $\mathrm{Tr}(M)$ is its trace. For two symmetric matrices A and B, we write $A \succcurlyeq B$ if $A - B$ is positive semidefinite. For a matrix M, $M_{j\cdot}$ and $M_{\cdot i}$ denote its j'th row and i'th column respectively. For a collection of indices J, M_{JJ} denotes the principal submatrix of M found by taking its columns and rows corresponding to indices in J, and for a vector x, $x[J]$ denotes the components of x corresponding to indices in J. For a matrix M, the operator $\mathrm{supp}(M)$ denotes its support, i.e. the indices corresponding to nonzero components in M. We denote the *reduced condition number* of Σ (with respect to the dimension k) as $\kappa := \frac{\lambda_1}{\lambda_k}$.

For two functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ if there exists a constant C such that $f(n) \leq Cg(n)$ for all n sufficiently large, and we write $f(n) \ll g(n)$ or $f(n) = o(g(n))$ if $f(n)/g(n) \to 0$ as $n \to \infty$. In the proofs, a generic constant C may change from line to line.

2 SPARSE PCA AND SPARSISTENCY

Suppose $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ are mean-zero random variables with covariance matrix Σ and eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Define the empirical covariance $\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$, which is just the usual method of moments estimator. We assume that Σ has a *sparse* k-dimensional leading subspace, meaning that its leading

k eigenvectors are s-sparse, in the sense that there is a set $J \subset \{1, \dots, p\}$ with cardinality at most s, with each eigenvector's nonzero support restricted to indices in J. In the language of Vu and Lei (2013), this setting refers to *row*-sparsity (as opposed to *column*-sparsity). See Vu and Lei (2013) for a comparison. We denote the $p \times k$ matrix U as the matrix of k orthonormal eigenvectors of Σ. Since U is assumed row-sparse, it has at most s nonzero *rows*. Concretely, this means that the nonzero support of each column of U is restricted to rows with indices in J. A useful interpretation of the set J is that it corresponds to the subset of covariates that contribute to the directions of maximum variance. In order for Σ to have a well-defined (sparse) leading k-dimensional subspace, it must have an eigengap, meaning that $\lambda_k - \lambda_{k+1} > 0$. In Section 3, Assumption 4 offers a slightly more quantitative condition on this eigengap.

The *sparse PCA problem* consists of estimating the matrix U from the observations $\{X_i\}_{i=1}^n$. There have been a number of approaches, including, but not limited to semidefinite programming Amini and Wainwright (2009); d'Aspremont et al. (2007), Fantope Projection and Selection algorithm (Vu et al., 2013; Lei and Vu, 2015), nonconvex approaches (Gu et al., 2014), Bayesian approaches (Xie et al., 2019), amongst others (Gataric et al., 2020; Chen and Rohe, 2020; Wang et al., 2014; Ma, 2013). In this paper we consider any algorithm that selects the correct support with high probability (see Assumption 2) in an asymptotic regime where $k \ll s \ll n \lesssim p$. From a practical standpoint, it is useful to consider the regime where k stays fixed but s tends to infinity as n and p at a rate $s = o(n)$. This regime is similar to that studied in the literature on high-dimensional sparse linear models, where one assumes that the coefficients are s-sparse with $s \ll n$. While it is possible to use analogous techniques to those in sparse linear models to study sparse PCA (e.g. Janková and van de Geer (2021)), the unsupervised problem of sparse PCA is markedly distinct from the *supervised* setting of sparse linear regression, and often requires additional considerations.

Note that if Π is a permutation matrix, then $\Pi\Sigma\Pi^\top(\Pi U) = \Pi\Sigma U = \Pi U \Lambda$, where Λ is the $k \times k$ diagonal matrix of leading eigenvalues of Σ. This shows that ΠU are eigenvectors of $\Pi\Sigma\Pi^\top$. Therefore, given the set of nonzero indices J, without loss of generality, we can assume $J = \{1, \dots, s\}$ by permuting Σ if necessary. We can partition Σ via

$$\Sigma := \begin{pmatrix} \Sigma_{JJ} & \Sigma_{JJ^c} \\ \Sigma_{JJ^c}^\top & \Sigma_{J^cJ^c} \end{pmatrix};$$

a similar partition holds for $\hat{\Sigma}$ and U. Under the assumption that the leading eigenvectors of Σ are sparse, we have from the eigenvector equation that

$$\Sigma U = \begin{pmatrix} \Sigma_{JJ} & \Sigma_{JJ^c} \\ \Sigma_{JJ^c}^\top & \Sigma_{J^cJ^c} \end{pmatrix} \begin{pmatrix} U_J \\ 0 \end{pmatrix} = \begin{pmatrix} \Sigma_{JJ} U_J \\ \Sigma_{JJ^c}^\top U_J \end{pmatrix} = \begin{pmatrix} U_J \\ 0 \end{pmatrix} \Lambda$$

which shows also that U_J is orthogonal to the matrix $\Sigma_{JJ^c}^\top$ and that the leading k eigenvectors and eigenvalues of Σ_{JJ} are exactly the leading k eigenvectors of Σ with the zeros removed.

Algorithm 1

```
Require: Sparsistent sparse PCA algorithm SparsePCA, empirical covariance matrix $\hat{\Sigma}$
1: Run SparsePCA algorithm on $\hat{\Sigma}$, obtaining support set estimate $\hat{J} \subset \{1, \dots, p\}$.
2: Define $\tilde{U}_{\hat{J}}$ as the leading $k$ eigenvectors of $\hat{\Sigma}_{\hat{J}\hat{J}}$.
   return Full matrix $\tilde{U}$, where
```

$$\tilde{U}_{i\cdot} = \begin{cases} (\tilde{U}_{\hat{J}})_{i\cdot} & i \in \hat{J} \\ 0 & i \notin \hat{J} \end{cases}$$

An important property of any sparse PCA algorithm is identifying the support J with high probability. Suppose $\hat{U}$ is any estimator for U (or, equivalently, $\widehat{UU^\top}$ is any estimator for $UU^\top$). In this work we consider a "debiased" version of sparse PCA under the assumption that $\hat{U}$ and U contain the same set of nonzero components, which implies that the estimator $\hat{U}$ equivalently estimates the support J. We defer the particular details of this assumption to Assumption 2. Our estimator is then defined as the following modification on any sparsistent algorithm: given any set J, let $\tilde{U}_J$ be the $s \times k$ matrix of eigenvectors of the principal submatrix $\hat{\Sigma}_{JJ}$, and define $\tilde{U} := \begin{pmatrix} \tilde{U}_J \\ 0 \end{pmatrix}$. If the algorithm is sparsistent, then the correct set J will be selected with high probability. In this way, the particular choice of sparse PCA algorithm can be viewed as a variable selection procedure as opposed to an estimation procedure. The full procedure is presented in Algorithm 1.

A natural question is whether sparsistent algorithms for sparse PCA exist. The answer is positive: in Theorem 1 of Lei and Vu (2015), the authors provide deterministic conditions on Σ guaranteeing that the Fantope Projection and Selection estimator is unique and has support set J with probability at least $1 - O(p^{-2})$ when $s\sqrt{\frac{\log(p)}{n}} \to 0$. Their conditions require an error bound on $\|\hat{\Sigma} - \Sigma\|_{\max}$ as well as conditions on the magnitudes of the eigengaps and entries of the projection matrices. Similarly, Gu et al. (2014) provide general conditions on Σ (in terms of the magnitudes of the entries) so that their (nonconvex) algorithm obtains the support set J with probability at least $1-O(n^{-2})$ when $\frac{sk\log(p)}{n} \to 0$. In general, sparsistency is a property of an algorithm, and the particular structure of Σ must be taken into account. Therefore, our results will hold for general matrices Σ with only mild conditions, and can be coupled with additional structural assumptions and algorithms to yield improved recovery guarantees.

3 MAIN RESULTS

In order to state our main results, we need a few assumptions. Our main results will be stated for large n with p, s and k functions of n. We have the following assumption on the dimensions.

Assumption 1 (Sample Size and Dimension). *The sample size n and dimension p satisfy*

$$s\log(p) \ll n; \qquad k \ll s.$$

The assumption that $s\log(p) \ll n$ is weaker than the assumption $s \lesssim \sqrt{n/\log(p)}$ as is the condition in Lei and Vu (2015) for sparsistency. However, this still allows $p/n \to \infty$; e.g. $p = n^c$ for any $c \geq 1$. The second condition $k \ll s$ is not explicitly required, but it does rule out the degenerate case $k = O(s)$, since $k \leq s$ by definition. In many works $k = 1$ (e.g. Amini and Wainwright (2009); Elsener and van de Geer (2019); Janková and van de Geer (2021)).

The next assumption imposes the condition that whatever variable selection procedure we use selects the correct support set J with high probability.

Assumption 2 (Sparsistency). *The algorithm is* sparsistent, *meaning that with probability $1-\delta$ the correct set J is chosen.*

Note that Theorem 1 of Lei and Vu (2015) provides sufficient conditions for Assumption 2 to hold, as does Theorem 1 of Gu et al. (2014). In general, this assumption is the hardest to check as it depends on the particular variable selection algorithm. In Lei and Vu (2015), the authors show that $\delta = O(p^{-2})$ when $s\sqrt{\frac{\log(p)}{n}} \to 0$ (in addition to some other conditions omitted here). Similarly, Gu et al. (2014) show that $\delta = O(n^{-2})$ when $\frac{s\log(p)}{n} \to 0$ (in addition to other conditions omitted here). Typically the other conditions include some "signal-strength" requirements, such as the magnitudes of the entries of Σ being sufficiently large. The particular details for these requirements can be found in Lei and Vu (2015) and Gu et al. (2014) respectively.

The following assumption imposes general tail conditions on the distribution of the observations $X_1, \ldots, X_n$.

Assumption 3 (Randomness). *The variables X_i are mean zero and satisfy $X_i = \Sigma^{1/2} Y_i$ for independent random variables Y_i with independent coordinates with unit variance. Furthermore, the ψ_2 norm of each coordinate Y_{ij} satisfies $\|Y_{ij}\|_{\psi_2} = 1$.*

This assumption says that the X_i's are linear combinations of Y_i's whose entries are independent. In general, assuming that each observation is a linear combination of independent random variables is a little stringent, but still common in the random matrix theory literature (e.g. El Karoui (2010); Knowles and Yin (2017); Bao et al. (2020); Ding (2021); Yang (2019, 2020)). While a more general result may be possible, Assumption 3 includes the setting that the Y_i's are i.i.d. Gaussians with identity covariance.

The following assumption imposes a quantitative condition on the eigengap (note that the existence of an eigengap is required for identifiability).

Assumption 4 (Eigenvalues). *The top eigenvalues of Σ satisfy*

$$C\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right) + \frac{\lambda_{k+1}}{8} \leq \frac{\lambda_k}{8}$$

for some sufficiently large constant C. In addition, for all p, we have that $2\lambda_{k+1} < (1-\varepsilon)\lambda_k$ for some $\varepsilon > \frac{1}{64}$.

The requirement $\varepsilon > \frac{1}{64}$ is somewhat arbitrary and can be relaxed in general to any constant strictly greater than zero. The other part of the assumption is required to obtain enough signal on the top k eigenvalues of Σ, and hence Σ_{JJ}. Furthermore, in light of Lemma 1 (our principal submatrix concentration bound), this ensures that the top k eigenvalues of $\hat{\Sigma}_{JJ}$ "track" those of Σ_{JJ}. In lieu of stronger assumptions, such as in a spiked model, this is the minimum requirement to guarantee that leading eigenvectors of $\hat{\Sigma}_{JJ}$ are well-defined.

The main results will be stated in terms of the $2 \to \infty$ norm of the difference of two matrices. Recall that for a matrix $M \in \mathbb{R}^{p\times k}$, we have that

$$\|M\|_{2\to\infty} = \max_{1\leq i\leq p} \|M_{i\cdot}\|_2;$$

that is, $\|M\|_{2\to\infty}$ is the maximum (Euclidean) row norm of the matrix M. Moreover, the $2 \to \infty$ norm has some attractive geometrical properties; for example, for two matrices A and B, we have that $\|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|$. More discussion on these relationships can be found in Cape et al. (2019b).

The following assumption concerns the *incoherence* of the matrix U, which is defined as $\|U\|_{2\to\infty}$. This assumption is only included to ease interpretation and is not explicitly required. A more general – albeit more complicated – result is provided in the supplementary material.

Assumption 5 (Incoherence and Conditioning)**.** *Suppose $\|U\|_{2\to\infty} \lesssim \left(\frac{k}{s}\right)^{1/2}$, that $k \lesssim \sqrt{s}$, and that the eigenvalues satisfy*

$$\lambda_{k+1} \leq \frac{\lambda}{2} < \lambda \leq \lambda_k \leq \lambda_1 \leq \kappa\lambda$$

for some parameters κ and λ.

The requirement $k \lesssim \sqrt{s}$ is only needed to simplify terms. The incoherence assumption states that the matrix Σ_{JJ} is incoherent in the usual sense. In this paper we do not worry about the particular incoherence constant as long as it is $O(1)$, whereas in the matrix completion literature (Candes and Plan, 2010; Candes and Tao, 2010; Chen et al., 2020b, 2019) one often studies the precise dependence on the incoherence constant. If one desires a more refined understanding of incoherence, our more general result in the supplementary material shows how our upper bound depends explicitly on the incoherence of U.

In addition, Assumption 5 should not be confused with Assumption 4 on the eigengap. The parameter κ is the *reduced condition number* of the leading k-dimensional subspace of Σ, and can be much smaller than the usual (full) condition number of Σ, especially when the leading k eigenvalues are of comparable order (or "spiked") relative to the bottom $p-k$ eigenvalues. Assumption 4 in fact implies an upper bound on κ of order at most $\sqrt{n/(s\log(p))}$, but it is useful to think of the setting that $\kappa = O(1)$, which corresponds to the case where the leading k eigenvalues are of comparable order. In the setting that the eigenvalues are uniformly bounded away from zero and infinity, this assumption is not particularly strong; moreover, if the leading k eigenvalues grow sufficiently fast as a function of n and p, then the leading k eigenvectors are consistent without additional assumptions. Consequently, the primary technical condition in Assumption 5 is on the incoherence, i.e. $\|U\|_{2\to\infty} \lesssim \left(\frac{k}{s}\right)^{1/2}$.

Before stating the main theorem, we will require some notions from subspace perturbation theory (Bhatia, 1997; Stewart and Sun, 1990). For $V, V' \in \mathbb{O}(p,k)$, the quantity

$$d_F(V,V') = \inf_{W\in\mathbb{O}(k)} \|V - V'W\|_F \qquad (1)$$

defines a metric on k-dimensional subspaces invariant to choice of basis. Therefore, by analogy, one might wish to study the quantity

$$d_{2\to\infty}(V,V') := \inf_{W\in\mathbb{O}(k)} \|V - V'W\|_{2\to\infty}. \qquad (2)$$

Unfortunately, for fixed V, V', one cannot necessarily compute the minimizer in (2) in closed form. However, for fixed V, V' the minimizer of (1) is attained using the singular value decomposition of $V^\top V'$. That is, let $W_1 D W_2^\top$ be the singular value decomposition of $V^\top V'$. Then the minimizer of (1), denoted W_*, satisfes $W_* := W_1 W_2^\top$. In addition,

$$d_{2\to\infty}(V,V') \leq \|V - V'W_*\|_{2\to\infty}.$$

Therefore, the results will be stated in terms of the *existence* of an orthogonal matrix $W_* \in \mathbb{O}(k)$ that provides an upper bound for the $2 \to \infty$ distance. In the proof, we show that W_* is actually a specific Frobenius-optimal orthogonal matrix. For convenience, we also include more information on subspace distances in the supplementary material (Appendix C).

We are now prepared to state our main result.

Theorem 1. *Suppose Assumptions 1, 2, 3, 4, and 5 are satisfied, and let $\tilde{U}$ be the output of Algorithm 1. Then with probability at least $1 - \delta - p^{-2}$, there exists an orthogonal matrix $W_* \in \mathbb{O}(k)$ such that*

$$\max_{1\leq i\leq n} \|\tilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \kappa^2 \sqrt{\frac{k\log(p)}{n}} + \kappa^3 \frac{s\log(p)}{n}.$$

Consequently, if $\kappa = O(1)$, then

$$\max_{1\leq i\leq n} \|\tilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \sqrt{\frac{k\log(p)}{n}} + \frac{s\log(p)}{n}.$$

As a brief remark, the dependence on the reduced condition number κ here may be suboptimal and could potentially be improved – we believe this is primarily an artifact of our proof technique and not a fundamental requirement. Recall that in the regime that the eigenvalues are bounded away from zero and infinity, when the leading k eigenvalues are of comparable order, it holds that $\kappa = O(1)$.

Note that by taking $\delta = O(p^{-2})$ and the conditions in Lei and Vu (2015) needed for sparsistency, the above bound holds with probability at least $1 - O(p^{-2})$; similarly, under the conditions needed for sparsistency in Gu et al. (2014), one has $\delta = O(n^{-2})$, in which case the bound holds with probability at least $1 - O(n^{-2})$.

4 DISCUSSION

In the regime that the eigenvalues are uniformly bounded away from zero and infinity in n, then Theorem 1 shows that we have the error rate

$$\max_{1\leq i\leq n} \|\tilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \max\left(\sqrt{\frac{k\log(p)}{n}}, \frac{s\log(p)}{n}\right).$$

In contrast, under the same conditions, in Frobenius norm, it has been shown in Cai et al. (2013) that the minimax rate satisfies

$$\|\tilde{U} - UW_*\|_F \lesssim \sqrt{\frac{s\log(p)}{n}},$$

so Theorem 1 improves upon this. Moreover, our result improves greatly upon the Frobenius norm bound in Vu et al. (2013), as well as the Frobenius minimax rates studied in Cai et al. (2013) and Vu and Lei (2013). To the best of our knowledge, this is the first $2 \to \infty$ guarantee for sparse PCA under a generic sparsistency requirement. A similar result was found in Xie et al. (2019) for spiked sparse covariance matrices, but here the only assumption on the spike is Assumption 4, which is a much weaker assumption.

Our bounds can also be compared to the spiked covariance matrix setting $\Sigma = U\Lambda U^\top + \sigma^2 I$, where U is no longer sparse but $\lambda_k \to \infty$ in n and p. In this setting the eigenvectors $\hat{U}$ of $\hat{\Sigma}$ are consistent in the following sense. Define the *effective rank* $\mathfrak{r}(\Sigma) := \frac{\mathrm{Tr}(\Sigma)}{\lambda_1}$. Theorem 1 of Cape et al. (2019b) (see also Yan et al. (2021) and Cai et al. (2021)) shows that if $\lambda_1 \gtrsim d/k$, $\mathfrak{r}(\Sigma) = o(n)$, $\kappa = O(1)$, and $\lambda_k - \sigma^2 \gtrsim \lambda_k$, then

$$\max_{1\le i\le n} \|\hat{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \sqrt{\frac{\max\{\mathfrak{r}(\Sigma), \log(d)\}}{n}}\sqrt{\frac{k^3}{p}}.$$

Here the primary error is no longer in *detecting* the leading eigenvectors (as the assumption that $\lambda_1 \gtrsim d/k$ implies large enough separation), but rather in the inherent statistical error implicit from the difference $\hat{\Sigma} - \Sigma$. Our upper bound requires that J is either known or can be estimated consistently (Assumption 2), so that our error depends on the inherent statistical error from $\hat{\Sigma}_{JJ} - \Sigma_{JJ}$. In contrast, we do not optimize for factors of λ_1 in our upper bound, as the setting for sparse PCA typically assumes that the eigenvalues remain bounded in n and p. We instead need only the (milder) eigenvalue separation in Assumption 4.

Suppose instead of just observing $X_1, \ldots, X_n \in \mathbb{R}^p$, one also observes response variables $Y_i \in \mathbb{R}$. Consider the linear model $Y_i = X_i^\top \beta + \varepsilon_i$, where ε_i is a mean-zero error term with variance σ^2. Suppose one first performs unsupervised dimensionality reduction on the data matrix via sparse PCA and then computes $\hat{\beta}$ using ordinary least squares with the reduced data matrix. The $2 \to \infty$ bound in Theorem 1 could provide a partial answer to the out-of-sample prediction performance using a variable selection procedure. To be concrete, define $\hat{\beta}$ as the output of ordinary least squares by regressing Y_i along $X\tilde{U}\tilde{U}^\top$, where $\tilde{U}$ is the output of the sparse PCA procedure in Algorithm 1 and X is the $n \times p$ matrix of predictors. Following Huang et al. (2020), we can bound the risk of a new sample point (x_*, Y_*) via

$$\begin{aligned}
&\mathbb{E}\|Y_* - x_*^\top \tilde{\beta}_{\mathrm{SPCA}}\|^2 | X \\
&\qquad \le \beta^\top (I - \tilde{U}\tilde{U}^\top)\Sigma(I - \tilde{U}\tilde{U}^\top)\beta \\
&\qquad\qquad + \frac{\sigma^2}{n}\mathrm{Tr}\left[\left(\frac{1}{n}\tilde{U}\tilde{U}^\top X^\top X \tilde{U}\tilde{U}^\top\right)^\dagger \Sigma\right] + \sigma^2,
\end{aligned}$$

where the first term represents the bias, the second term represents the variance, and the third term (σ^2) is the noise intrinsic to the problem. The bias term can be expanded further via

$$\begin{aligned}
&\beta^\top (I - \tilde{U}\tilde{U}^\top)\Sigma(I - \tilde{U}\tilde{U}^\top)\beta \\
&\quad = \beta^\top (\tilde{U}\tilde{U}^\top - UU^\top)\Sigma(\tilde{U}\tilde{U}^\top - UU^\top)\beta \\
&\qquad + 2\beta^\top (\tilde{U}\tilde{U}^\top - UU^\top)\Sigma(I - UU^\top)\beta \\
&\qquad + \lambda_{k+1}\|\beta\|_2^2.
\end{aligned}$$

Consider the second term. This could be bounded by noting that

$$\begin{aligned}
&|\beta^\top (\tilde{U}\tilde{U}^\top - UU^\top)\Sigma(I - UU^\top)\beta| \\
&\quad \le \left\|\beta^\top (\tilde{U}\tilde{U}^\top - UU^\top)\right\|_\infty \left\|\Sigma(I - UU^\top)\beta\right\|_1 \\
&\quad \le \lambda_{k+1}\|\beta\|_\infty \|\beta\|_1 \|\tilde{U}\tilde{U}^\top - UU^\top\|_{2\to\infty}.
\end{aligned}$$

This bound has a factor of $\|\tilde{U}\tilde{U}^\top - UU^\top\|_{2\to\infty}$, which, while not exactly the same as what appears in Theorem 1, is closely related to it by appealing to notions in subspace perturbation theory (see, e.g. Lemma 1 of Cai and Zhang (2018)). Therefore, through similar analysis, one could obtain bounds for the other bias and variance terms with respect to the eigenvalues of Σ, the quantity $\|\tilde{U}\tilde{U}^\top - UU^\top\|_{2\to\infty}$ and the quantities $\|\beta\|_1$ and $\|\beta\|_\infty$. Consequently, these bounds would complement those in Theorem 1 of Huang et al. (2020) as sparse PCA is typically needed in a regime when $\mathfrak{r}(\Sigma) \gtrsim n$, whereas Huang et al. (2020) study the setting that $\mathfrak{r}(\Sigma) = o(n)$.

Finally, our upper bound depends on the debiased estimator $\tilde{U}_J$, which is the matrix of eigenvectors of $\hat{\Sigma}_{JJ}$. A key requirement is that any algorithm obtains the correct set J with probability at least $1 - \delta$. In general, one must consider the output of an optimization procedure to determine whether a specific algorithm obtains the correct set J. If one additionally wanted to *test* whether a certain row of U is equal to zero (i.e., whether $i \in J$), then one would need to construct a different debiased estimator as in Janková and van de Geer (2021) that uses the first-order necessary optimality conditions. This procedure therefore relies heavily on the particular algorithm used, whereas our bounds hold for generic algorithms.

5 OVERVIEW OF THE PROOF OF THEOREM 1

The full proof of Theorem 1 is in the supplementary material, though we include a brief overview here. First, our main upper bound holds without Assumption 5, and we provide this general upper bound in Theorem 2 (stated in the supplementary material A) and show how Theorem 1 can be deduced from Assumption 5. To prove Theorem 2, we first show the following *principal submatrix concentration* bound.

Lemma 1 (Principal Submatrix Concentration)**.** *Let J be an index set of $\{1, ..., p\}$ of size s. Then*

$$\|\hat{\Sigma}_{JJ} - \Sigma_{JJ}\| \lesssim \lambda_1 \left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right)$$

with probability at least $1 - O(p^{-4})$.

The proof is somewhat standard and primarily follows arguments detailed in Wainwright (2019) via ε-nets and concentration, though we include it in Section B.1 for completeness. It is also very similar to a result in Amini and Wainwright (2009) for Gaussian random variables. To the best of our knowledge, there is no general result of this form in the literature for subgaussian random vectors. The following Lemma shows that the leading k eigenvalues of $\hat{\Sigma}_{JJ}$ are well-separated from its bottom eigenvalues.

Lemma 2 (Existence of an Eigengap)**.** *Under the event in Lemma 1 and Assumption 4, the eigenvalues of $\hat{\Sigma}_{JJ}$ and Σ_{JJ} satisfy*

$$\lambda_k - \tilde{\lambda}_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8}; \qquad \tilde{\lambda}_k - \lambda_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8};$$
$$\tilde{\lambda}_k \geq \frac{\lambda_k}{4}.$$

Consequently, this bound holds with probability at least $1 - O(p^{-4})$.

We also note that $\lambda_{k+1}(\Sigma_{JJ}) \leq \lambda_{k+1}$ by the Cauchy interlacing inequalities (Horn and Johnson, 2012), and the top k eigenvalues of Σ_{JJ} are the same as those of Σ by the eigenvector equation. These lemmas set the stage for our main analysis.

As an immediate consequence of Lemmas 1 and 2, we can obtain the following proposition concerning the spectral proximity of $U_J U_J^\top$ to $\tilde{U}_J \tilde{U}_J^\top$, ensuring that $\tilde{U}_J$ (and hence $\tilde{U}$) is well-defined.

Proposition 1 (Spectral Proximity)**.** *Under the assumptions of Theorem 2, we have that*

$$\|U_J U_J^\top - \tilde{U}_J \tilde{U}_J^\top\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}} \left[\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right]$$

with probability at least $1 - O(p^{-4})$.

We use this bound several times in our subsequent analysis. After these preliminary bounds, which are restated for convenience in the supplementary material, we develop an expansion for the difference $\tilde{U}_J - U_J W_*$ in terms of the error matrix $(\Sigma - \hat{\Sigma})$ and deterministic quantities depending only on Σ. Informally, we show that we have the "first-order" approximation

$$\tilde{U}_J - U_J W_* = (\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ}) \tilde{U}_J \tilde{\Lambda}^{-1} + R,$$

where R is a residual term and $\tilde{\Lambda}$ is the diagonal matrix of the k leading eigenvalues of $\hat{\Sigma}_{JJ}$. Lemma 2 ensures that the eigenvalues of $\tilde{\Lambda}$ can be bounded with respect to the eigenvalues of Σ. The residual term R (the terms T_1, T_2, and T_3 in the supplementary material) is bounded in Lemmas 3, 4, and 5 with tools from complex analysis (Greene and Krantz, 2006), matrix perturbation theory (Bhatia, 1997), and high-dimensional probability (Wainwright, 2019; Vershynin, 2018).

To bound the leading term in 2 → ∞ norm, we show that it can be further decomposed into two terms, that we dub J_1 and J_2, by the decomposition

$$(\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ}) \tilde{U}_J = (\hat{\Sigma}_{JJ} - \Sigma_{JJ}) \tilde{U}_J + U_\perp \Lambda_\perp U_\perp^\top \tilde{U}_J$$
$$:= J_1 + J_2,$$

where $U_\perp$ is the $s \times (s-k)$ matrix such that $[U_J, U_\perp]$ is an orthogonal matrix. The first term reflects the error from the randomness and the leading subspace U_J and the second term reflects the influence of $U_\perp$ on $\tilde{U}_J$.

The term $J_2 = U_\perp \Lambda_\perp U_\perp^\top \tilde{U}_J$ is bounded using a matrix series expansion for the matrix $\tilde{U}_J$ (Lemma 6). More explicitly, we define the perturbation $E := \hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top$, and we show that we can write

$$\tilde{U}_J = \sum_{m=0}^{\infty} E^m (U_J \Lambda U_J^\top) \tilde{U}_J \Lambda^{-m+1}.$$

We then analyze each term in 2 → ∞ norm, take a union bound for the first $O(\log(n))$ terms and bound the remaining part of the series coarsely using the spectral norm. Similar techniques have been used in Cape et al. (2019a); Xie et al. (2019); Tang (2018) and Tang et al. (2017), but our analysis requires additional considerations due to the fact that we do not have a mean-zero perturbation since $\mathbb{E}E = U_\perp U_\perp^\top \Sigma_{JJ} U_\perp U_\perp^\top$. However, the matrix EU_J *is* mean-zero since $U_\perp^\top U_J = 0$.

The remaining term $J_1 = (\hat{\Sigma}_{JJ} - \Sigma_{JJ}) \tilde{U}_J$ is then analyzed directly through its block-structure (Equation (7)). Letting X be the $n \times p$ matrix whose rows are the observations, by Assumption 3, $X = Y \Sigma^{1/2}$, where Y is an $n \times p$ matrix of independent random variables with unit variance. Then the empirical covariance

$\hat{\Sigma} = \frac{1}{n} X^\top X$ and hence

$$\hat{\Sigma}_{JJ} = \frac{1}{n}\Big((\Sigma^{1/2})_{JJ} Y_J^\top Y_J (\Sigma^{1/2})_{JJ} + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} + (\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top \Big)$$

where we have abused the notation

$$\Sigma^{1/2}_{JJ^c} = (\Sigma^{1/2})_{JJ^c}.$$

Above, the $n \times p$ matrix Y is partitioned via $Y = [Y_J, Y_{J^c}]$, where Y_J and Y_{J^c} are the variables corresponding to J and its complement, J^c, respectively. This term is bounded in Lemmas 7, 8, and 9. Lemmas 7 and 8 are standard applications of matrix perturbation theory (via Proposition 1) and standard concentration inequalities such as Bernstein's inequality, but Lemma 9 requires studying the spectral properties of the matrix Σ_{JJ^c} and its relation to U_J (Proposition 2).

Our proof is then completed by combining and aggregating all of these bounds. Throughout the proof we make heavy use of several important concentration inequalities and notions from subspace perturbation theory, so Appendix C in the supplementary material contains additional information on these topics.

Acknowledgements

Both authors thank the anonymous reviewers for feedback during the review process. JS's research was partially supported by NSF grant CCF 2007649. JA's research was partially supported by a fellowship from the Johns Hopkins Mathematical Institute of Data Science (MINDS) via its NSF TRIPODS award CCF-1934979, and through the Charles and Catherine Counselman Fellowship.

References

E. Abbe, J. Fan, and K. Wang. An ℓ_p theory of PCA and spectral clustering. *arXiv:2006.14062 [cs, math, stat]*, June 2020a.

E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3):1452–1474, June 2020b. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1854.

J. Agterberg, Z. Lubberts, and C. Priebe. Entrywise Estimation of Singular Vectors of Low-Rank Matrices with Heteroskedasticity and Dependence. *arXiv:2105.13346 [math, stat]*, May 2021.

A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, 37(5B):2877–2921, October 2009. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS664.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, July 2003. ISBN 978-0-471-36091-9. Google-Books-ID: Cmm9QgAACAAJ.

J. Banks, C. Moore, R. Vershynin, N. Verzelen, and J. Xu. Information-Theoretic Bounds and Phase Transitions in Clustering, Sparse PCA, and Submatrix Localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, July 2018. ISSN 1557-9654. doi: 10.1109/TIT.2018.2810020.

Z. Bao, X. Ding, J. Wang, and K. Wang. Statistical inference for principal components of spiked covariance matrices. *arXiv:2008.11903 [math, stat]*, September 2020.

Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(4):1780–1815, August 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1127.

R. Bhatia. *Matrix Analysis*, volume 169. Springer, 1997. ISBN 0-387-94846-5.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-20192-9.

C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49(2):944–967, April 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1986.

T. T. Cai and A. Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Annals of Statistics*, 46 (1):60–89, February 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1541.

T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *Annals of Statistics*, 41(6):3074–3110, December 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1178.

T. T. Cai, X. Han, and G. Pan. Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Annals of Statistics*, 48(3):1255–1280, June 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1798.

E. J. Candes and Y. Plan. Matrix Completion With Noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010. ISSN 0018-9219. doi: 10.1109/JPROC.2009.2035722.

E. J. Candes and T. Tao. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE*

Transactions on Information Theory, 56(5):2053–2080, May 2010. ISSN 1557-9654. doi: 10.1109/TIT.2010.2044061.

J. Cape, M. Tang, and C. E. Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250, March 2019a. ISSN 0006-3444. doi: 10.1093/biomet/asy070.

J. Cape, M. Tang, and C. E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Annals of Statistics*, 47(5):2405–2439, October 2019b. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1752.

V. Charisopoulos, A. R. Benson, and A. Damle. Entrywise convergence of iterative methods for eigenproblems. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5644–5655. Curran Associates, Inc., 2020.

F. Chen and K. Rohe. A New Basis for Sparse PCA. *arXiv:2007.00596 [cs, stat]*, July 2020.

Y. Chen, J. Fan, C. Ma, and Y. Yan. Inference and Uncertainty Quantification for Noisy Matrix Completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, November 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1910053116.

Y. Chen, Y. Chi, J. Fan, and C. Ma. Spectral Methods for Data Science: A Statistical Perspective. *arXiv:2012.08496 [cs, eess, math, stat]*, December 2020a.

Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization. *SIAM Journal on Optimization*, 30(4): 3098–3121, January 2020b. ISSN 1052-6234. doi: 10.1137/19M1290000.

A. Damle and Y. Sun. Uniform Bounds for Invariant Subspace Perturbations. *SIAM Journal on Matrix Analysis and Applications*, 41(3):1208–1236, January 2020. ISSN 0895-4798. doi: 10.1137/19M1262760.

A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, 49(3):434–448, January 2007. ISSN 0036-1445. doi: 10.1137/050645506.

X. Ding. Spiked sample covariance matrices with possibly multiple bulk components. *Random Matrices: Theory and Applications*, 10(01):2150014, January 2021. ISSN 2010-3263. doi: 10.1142/S2010326321500143.

N. El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, February 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS648.

A. Elsener and S. van de Geer. Sparse spectral estimation with missing and corrupted measurements. *Stat*, 8(1):e229, 2019. ISSN 2049-1573. doi: 10.1002/sta4.229.

J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96 (456):1348–1360, December 2001. ISSN 0162-1459. doi: 10.1198/016214501753382273.

J. Fan, W. Wang, and Y. Zhong. An ℓ_{∞} Eigenvector Perturbation Bound and Its Application. *Journal of Machine Learning Research*, 18 (207):1–42, 2018. ISSN 1533-7928.

J. Fan, Y. Fan, X. Han, and J. Lv. Asymptotic Theory of Eigenvectors for Random Matrices With Diverging Spikes. *Journal of the American Statistical Association*, 0(0):1–14, October 2020. ISSN 0162-1459. doi: 10.1080/01621459.2020.1840990.

C. Gao, Z. Ma, and H. H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, October 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1519.

M. Gataric, T. Wang, and R. J. Samworth. Sparse principal component analysis via axis-aligned random projections. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):329–359, 2020. ISSN 1467-9868. doi: 10.1111/rssb.12360.

R. E. Greene and S. G. Krantz. *Function Theory of One Complex Variable*. American Mathematical Soc., 2006. ISBN 978-0-8218-3962-1. Google-Books-ID: u5vhseYCcqkC.

Q. Gu, Z. Wang, and H. Liu. Sparse PCA with Oracle Property. *Advances in neural information processing systems*, 2014:1529–1537, 2014. ISSN 1049-5258.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

N. Huang, D. W. Hogg, and S. Villar. Dimensionality reduction, regularization, and generalization in overparameterized regressions. *arXiv:2011.11477 [cs, stat]*, November 2020.

J. Janková and S. van de Geer. De-Biased Sparse PCA: Inference for Eigenstructure of Large Covariance Matrices. *IEEE Transactions on Information Theory*, 67(4):2507–2527, April 2021. ISSN 1557-9654. doi: 10.1109/TIT.2021.3059765.

J. Jin, Z. T. Ke, and S. Luo. Estimating network memberships by simplex vertex hunting. *arXiv:1708.07852 [stat]*, September 2019.

I. M. Johnstone and A. Y. Lu. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association*, 104(486):682–693, June 2009. ISSN 0162-1459. doi: 10.1198/jasa.2009.0121.

A. Knowles and J. Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352, October 2017. ISSN 1432-2064. doi: 10.1007/s00440-016-0730-4.

V. Koltchinskii and D. Xia. Perturbation of Linear Forms of Singular Vectors Under Gaussian Noise. In C. Houdré, D. M. Mason, P. Reynaud-Bouret, and J. Rosiński, editors, *High Dimensional Probability VII*, Progress in Probability, pages 397–423, Cham, 2016. Springer International Publishing. ISBN 978-3-319-40519-3. doi: 10.1007/978-3-319-40519-3_18.

R. Krauthgamer, B. Nadler, and D. Vilenchik. Do semidefinite relaxations solve sparse PCA up to the information limit? *Annals of Statistics*, 43(3):1300–1322, June 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1310.

J. Lei and V. Q. Vu. Sparsistency and agnostic inference in sparse PCA. *Annals of Statistics*, 43(1):299–322, February 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1273.

L. Lei. Unified $\ell_{2\rightarrow\infty}$ Eigenspace Perturbation Theory for Symmetric Random Matrices. *arXiv:1909.04798 [math, stat]*, September 2019.

Z. Ma. Sparse principal component analysis and iterative thresholding. *Annals of Statistics*, 41(2):772–801, April 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1097.

X. Mao, P. Sarkar, and D. Chakrabarti. Estimating Mixed Memberships With Sharp Eigenvector Deviations. *Journal of the American Statistical Association*, 0(0):1–13, April 2020. ISSN 0162-1459. doi: 10.1080/01621459.2020.1751645.

K. Rohe and M. Zeng. Vintage Factor Analysis with Varimax Performs Statistical Inference. *arXiv:2004.05387 [math, stat]*, April 2020.

G. W. Stewart and J.-G. Sun. *Matrix pertubation theory*. Academic Press, 1990.

M. Tang. The eigenvalues of stochastic blockmodel graphs. *arXiv:1803.11551 [cs, stat]*, March 2018.

M. Tang, J. Cape, and C. E. Priebe. Asymptotically efficient estimators for stochastic blockmodels: the naive MLE, the rank-constrained MLE, and the spectral. *arXiv:1710.10936 [stat]*, October 2017.

R. 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.

V. Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, December 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1151.

V. Q. Vu, J. Cho, J. Lei, and K. Rohe. Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2670–2678. Curran Associates, Inc., 2013.

M. Wahl. A note on the prediction error of principal component regression. *arXiv:1811.02998 [math, stat]*, April 2019a.

M. Wahl. On the perturbation series for eigenvalues and eigenprojections. *arXiv:1910.08460 [math, stat]*, October 2019b.

M. J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using $_1$ -Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009. doi: 10.1109/TIT.2009.2016018.

M. J. Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint, February 2019. ISBN: 9781108627771 9781108498029 Library Catalog: www.cambridge.org Publisher: Cambridge University Press.

T. Wang, Q. Berthet, and R. J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44 (5):1896–1930, October 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1369.

Z. Wang, H. Lu, and H. Liu. Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time. *Advances in neural information processing systems*, 2014:3383–3391, 2014. ISSN 1049-5258.

D. Xia. Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15(2):3798–3851, January 2021. ISSN 1935-7524, 1935-7524. doi: 10.1214/21-EJS1876.

D. Xia and M. Yuan. Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a), November 2020. ISSN 1467-9868. doi: https://doi.org/10.1111/rssb.12400.

F. Xie. Entrywise limit theorems of eigenvectors and their one-step refinement for sparse random graphs. *arXiv:2106.09840 [math, stat]*, June 2021.

F. Xie, Y. Xu, C. E. Priebe, and J. Cape. Bayesian Estimation of Sparse Spiked Covariance Matrices in

High Dimensions. *arXiv:1808.07433 [math, stat]*, January 2019.

Y. Yan, Y. Chen, and J. Fan. Inference for Heteroskedastic PCA with Missing Data. *arXiv:2107.12365 [cs, math, stat]*, July 2021.

F. Yang. Edge universality of separable covariance matrices. *Electronic Journal of Probability*, 24, 2019. ISSN 1083-6489. doi: 10.1214/19-EJP381.

F. Yang. Linear spectral statistics of eigenvectors of anisotropic sample covariance matrices. *arXiv:2005.00999 [math, stat]*, May 2020.

Z. Yang, Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang. Sparse Nonlinear Regression: Parameter Estimation and Asymptotic Inference. *arXiv:1511.04514 [cs, math, stat]*, November 2015.

Y. Yu, T. Wang, and R. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102, April 2014. doi: 10.1093/biomet/asv008.

P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7 (90):2541–2563, 2006. ISSN 1533-7928.

Supplementary Material: Entrywise Recovery Guarantees for Sparse PCA via Sparsistent Algorithms

Abstract

This supplementary material contains all the proofs of our main results. Appendix A contains the proof of Theorem 1, Appendix B contains the proofs of additional lemmas needed en route the the proof of the main theorem, and Appendix C contains additional background material on Orlicz norms, concentration inequalities, and subspace perturbation theory.

A Proof of Theorem 1

In this section we prove Theorem 1. First, Theorem 1 is actually a consequence of the following more general theorem that does not require Assumption 5. Section A.1 develops the preliminary bounds in terms of principal submatrix and eigenvalue concentration (Lemmas 1 and 2), and in Section A.2 we prove Theorem 2. In Section A.3 we show how Theorem 1 can be deduced by combining Theorem 2 with Assumption 5. En route to the proof of Theorem 2 we introduce several technical lemmas; we prove these in Section B. Recall that we denote $\kappa := \frac{\lambda_1}{\lambda_k}$ as the (reduced) condition number of Σ.

Theorem 2. *Suppose Assumptions 1, 2, 3, and 4 are satisfied. Then with probability at least $1-\delta-p^{-2}$, there exists an orthogonal matrix $W_* \in \mathbb{O}(k)$ such that*

$$\max_{1\leq i\leq p} \|\tilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 + \mathcal{E}_5$$

where

$$\begin{aligned}
\mathcal{E}_1 &:= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}} \frac{s\log(p)}{n}\|U\|_{2\to\infty} + \kappa k\sqrt{\frac{\log(p)}{n}}\|U\|_{2\to\infty} \\
\mathcal{E}_2 &:= \frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s\log(p)}{n}\|U\|_{2\to\infty} \\
\mathcal{E}_3 &:= \sqrt{\frac{s\log(p)}{n}} \frac{\kappa\lambda_1^{1/2}}{\lambda_k - \lambda_{k+1}} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty}\right) \\
\mathcal{E}_4 &:= \frac{\lambda_{k+1}}{\lambda_k}\kappa^2\sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_{k+1}}{\lambda_k}\kappa^3\frac{s\log(p)}{n}; \\
\mathcal{E}_5 &:= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}} \frac{s\log(p)}{n} + \kappa\sqrt{\frac{k\log(p)}{n}}.
\end{aligned}$$

A.1 Preliminary Bounds

Note that by Assumption 2, we need only examine the $s \times k$ matrix of eigenvectors of $\hat{\Sigma}_{JJ}$ and Σ_{JJ} respectively. We will develop an expansion for the difference $\tilde{U}_J - U_J W_*$ by viewing $\hat{\Sigma}_{JJ}$ as a perturbation of Σ_{JJ}. For convenience we restate the initial preliminary bounds in the main paper. Except for Proposition 1, the proofs are in Section B.1. The first is the following principal submatrix concentration bound.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

Lemma 1 (Principal Submatrix Concentration)**.** *Let J be an index set of $\{1, ..., p\}$ of size s. Then*

$$\|\hat{\Sigma}_{JJ} - \Sigma_{JJ}\| \lesssim \lambda_1 \left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right)$$

with probability at least $1 - O(p^{-4})$.

Henceforth, we assume the correct support set J is known; it is the correct set J with probability at least $1 - \delta$ by Assumption 2. As discussed in the main paper, using Lemma 1, we can derive the following eigenvalue bound, which we present as a lemma below.

Lemma 2 (Existence of an Eigengap)**.** *Under the event in Lemma 1 and Assumption 4, the eigenvalues of $\hat{\Sigma}_{JJ}$ and Σ_{JJ} satisfy*

$$\lambda_k - \tilde{\lambda}_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8}; \qquad \tilde{\lambda}_k - \lambda_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8};$$
$$\tilde{\lambda}_k \geq \frac{\lambda_k}{4}.$$

Consequently, this bound holds with probability at least $1 - O(p^{-4})$.

Finally, we have the following $\sin \Theta$ distance error between U_J and $\tilde{U}_J$.

Proposition 1 (Spectral Proximity)**.** *Under the assumptions of Theorem 2, we have that*

$$\|U_J U_J^\top - \tilde{U}_J \tilde{U}_J^\top\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}} \left[\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right]$$

with probability at least $1 - O(p^{-4})$.

Proof of Proposition 1. By the Davis-Kahan Theorem (Bhatia, 1997; Yu et al., 2014) and Lemma 2,

$$\begin{aligned} \|U_J U_J^\top - \tilde{U}_J \tilde{U}_J^\top\| &\leq \frac{\|\hat{\Sigma}_{JJ} - \Sigma_{JJ}\|}{\lambda_k - \tilde{\lambda}_{k+1}}. \\ &\lesssim \frac{\|\hat{\Sigma}_{JJ} - \Sigma_{JJ}\|}{\lambda_k - \lambda_{k+1}} \end{aligned} \tag{3}$$

By Lemma 1, with probability at least $1 - O(p^{-4})$, the numerator can be bounded by

$$\|\hat{\Sigma}_{JJ} - \Sigma_{JJ}\| \leq \lambda_1 \left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right)$$

Combining this and Equation (3) gives the result. □

In the proofs that follow, we will use the fact that by Proposition 1, we have that

$$\|U_J U_J^\top - \tilde{U}_J \tilde{U}_J^\top\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}} \sqrt{\frac{s \log(p)}{n}},$$

which is a little more amenable to downstream analysis. In addition, we use several equivalent expressions for the spectral norm of the difference of projections; see Lemma 10 in Appendix C for a discussion of how to equate these.

A.2 Proof of Theorem 2

We now proceed with the proof. At a high level, the argument consists of a deterministic matrix decomposition, each term of which we bound in probability. Define $\tilde{\Lambda}$ as the diagonal $k \times k$ matrix of eigenvalues of $\hat{\Sigma}_{JJ}$. Define W_* to be the matrix

$$W_* := \underset{W \in \mathbb{O}(k)}{\arg\min} \|\tilde{U}_J - U_J W\|_F.$$

Is is well-known that W_* can be computed from the singular value decomposition of $U_J^\top \tilde{U}_J$ (e.g. Abbe et al. (2020b); Cape et al. (2019b); Chen et al. (2020a)).

We now expand the difference via

$$\begin{aligned} \tilde{U}_J - U_J W_* &= \tilde{U}_J - U_J U_J^\top \tilde{U}_J - U_J(W_* - U_J^\top \tilde{U}_J) \\ &= \tilde{U}_J - U_J \Lambda U_J^\top \tilde{U}_J \tilde{\Lambda}^{-1} + U_J \Lambda U_J^\top \tilde{U}_J \tilde{\Lambda}^{-1} - U_J U_J^\top \tilde{U}_J - U_J(W_* - U_J^\top \tilde{U}_J) \\ &= \tilde{U}_J - U_J \Lambda U_J^\top \tilde{U}_J \tilde{\Lambda}^{-1} + U_J(\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J \tilde{\Lambda})\tilde{\Lambda}^{-1} - U_J(W_* - U_J^\top \tilde{U}_J) \\ &= \tilde{U}_J \tilde{\Lambda}\tilde{\Lambda}^{-1} - U_J \Lambda U_J^\top \tilde{U}_J \tilde{\Lambda}^{-1} + U_J(\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J \tilde{\Lambda})\tilde{\Lambda}^{-1} - U_J(W_* - U_J^\top \tilde{U}_J) \\ &= (\tilde{U}_J \tilde{\Lambda} - U_J \Lambda U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1} + U_J(\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J \tilde{\Lambda})\tilde{\Lambda}^{-1} - U_J(W_* - U_J^\top \tilde{U}_J) \\ &= A + T_1 - T_2, \end{aligned} \tag{4}$$

where

$$\begin{aligned} A &:= (\tilde{U}_J \tilde{\Lambda} - U_J \Lambda U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1}; \\ T_1 &:= U_J(\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J \tilde{\Lambda})\tilde{\Lambda}^{-1} \\ T_2 &:= U_J(W_* - U_J^\top \tilde{U}_J). \end{aligned}$$

Both T_1 and T_2 are analyzed concisely in Lemmas 3 and 4 as follows. Their proofs are in Section B.2. The proof of Lemmas 3 and 4 are both rather straightforward and based on previous results in entrywise eigenvector analysis (Abbe et al., 2020a,b; Agterberg et al., 2021; Cai et al., 2021; Cape et al., 2019a,b; Chen et al., 2020a; Tang et al., 2017; Xia and Yuan, 2020; Xie et al., 2019; Xie, 2021; Yan et al., 2021).

Lemma 3 (Bound on T_1). *We have that*

$$\begin{aligned} \|U_J(\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J \tilde{\Lambda})\tilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})}\frac{s\log(p)}{n} \\ &\quad + \frac{k\lambda_1\|U\|_{2\to\infty}}{\lambda_k}\sqrt{\frac{\log(p)}{n}} \\ &\equiv \mathcal{E}_1 \end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Lemma 4 (Bound on T_2). *We have that*

$$\begin{aligned} \|U_J(W_* - U_J^\top \tilde{U}_J)\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2}\frac{s\log(p)}{n} \\ &\equiv \mathcal{E}_2 \end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Expanding Equation (4) into T_3 and T_4:

We further expand the remaining term in (4) by viewing $\hat{\Sigma}_{JJ}$ as a perturbation of $U_J U_J^\top \Sigma_{JJ}$ and using the eigenvector-eigenvalue equation via

$$\begin{aligned} A &= (\tilde{U}_J \tilde{\Lambda} - U_J \Lambda U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1} \\ &= (\hat{\Sigma}_{JJ}\tilde{U}_J - \Sigma_{JJ} U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1} \\ &= (U_J U_J^\top \Sigma_{JJ}\tilde{U}_J + (\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\tilde{U}_J - \Sigma_{JJ} U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1} \\ &= U_J U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1} + (\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\tilde{U}_J \tilde{\Lambda}^{-1} \\ &= T_3 + T_4, \end{aligned}$$

where

$$\begin{aligned} T_3 &:= U_J U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1} \\ T_4 &:= (\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\tilde{U}_J \tilde{\Lambda}^{-1}. \end{aligned}$$

The term T_3 can be analyzed via techniques from complex analysis. We present this bound as a lemma, but defer the proof to Section B.3.

Lemma 5 (Bound on T_3)**.** *We have that*

$$\begin{aligned}\|U_J U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \sqrt{\frac{s\log(p)}{n}}\frac{\lambda_1^{3/2}}{\lambda_k(\lambda_k-\lambda_{k+1})}\min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty}\right)\\ &\equiv \mathcal{E}_3\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Expanding T_4 in terms of J_1 and J_2:

We now proceed to bound T_4. We have by Lemma 2 and properties of the $2 \to \infty$ norm that

$$\begin{aligned}\|(\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\tilde{U}_J\tilde{\Lambda}^{-1}\|_{2\to\infty} &\leq \frac{1}{\tilde{\lambda}_k}\|(\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\tilde{U}_J\|_{2\to\infty}\\ &\lesssim \frac{1}{\lambda_k}\|(\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\tilde{U}_J\|_{2\to\infty}. \qquad (5)\end{aligned}$$

Note that $\tilde{U}_J$ is the matrix of eigenvectors of $\hat{\Sigma}_{JJ}$ and hence is not independent of the error matrix $\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ}$, so one cannot bound the maximum row norm of the matrix above with standard concentration techniques. Let $U_\perp$ be the matrix such that $[U_J, U_\perp]$ is an $s \times s$ orthogonal matrix, and let $\tilde{U}_\perp$ be defined similarly. Define also $\Lambda_\perp$ and $\tilde{\Lambda}_\perp$ as the matrix of bottom $s-k$ eigenvalues of Σ_{JJ} and $\hat{\Sigma}_{JJ}$ respectively. Since $\tilde{U}_\perp^\top \tilde{U}_J = 0$, we have that

$$\begin{aligned}\frac{1}{\lambda_k}\|(\hat{\Sigma}_{JJ} - &U_J U_J^\top \Sigma_{JJ})\tilde{U}_J\|_{2\to\infty}\\ &= \frac{1}{\lambda_k}\left\|\left(\tilde{U}_J\tilde{\Lambda}\tilde{U}_J^\top + \tilde{U}_\perp\tilde{\Lambda}_\perp\tilde{U}_\perp^\top - U_J\Lambda_J U_J^\top\right)\tilde{U}_J\right\|_{2\to\infty}\\ &\leq \frac{1}{\lambda_k}\left\|\left(\tilde{U}_J\tilde{\Lambda}\tilde{U}_J^\top + \tilde{U}_\perp\tilde{\Lambda}_\perp\tilde{U}_\perp^\top - U_J\Lambda_J U_J^\top - U_\perp\Lambda_\perp U_\perp^\top\right)\tilde{U}_J\right\|_{2\to\infty} + \frac{1}{\lambda_k}\|U_\perp\Lambda_\perp U_\perp^\top\tilde{U}_J\|_{2\to\infty}\\ &\leq \frac{1}{\lambda_k}\|(\hat{\Sigma}_{JJ} - \Sigma_{JJ})\tilde{U}_J\|_{2\to\infty} + \frac{1}{\lambda_k}\|U_\perp\Lambda_\perp U_\perp^\top\tilde{U}_J\|_{2\to\infty}\\ &:= \frac{1}{\lambda_k}\|J_1\|_{2\to\infty} + \frac{1}{\lambda_k}\|J_2\|_{2\to\infty} \qquad (6)\end{aligned}$$

where

$$\begin{aligned}J_1 &:= (\hat{\Sigma}_{JJ} - \Sigma_{JJ})\tilde{U}_J;\\ J_2 &:= U_\perp\Lambda_\perp U_\perp^\top\tilde{U}_J.\end{aligned}$$

The term J_2 can be bounded in the following lemma, but it is rather technical; moreover, it requires some analysis that is relatively novel in the subspace estimation literature; in particular, we combine some ideas from Xia and Yuan (2020) as well as Cape et al. (2019a); Xie et al. (2019); Tang (2018); Tang et al. (2017). The proof is in Section B.4.

Lemma 6 (Bound on J_2)**.** *The term J_2 satisfies*

$$\begin{aligned}\|U_\perp\Lambda_\perp U_\perp^\top\tilde{U}_J\|_{2\to\infty} &\lesssim \kappa^2\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}} + \lambda_{k+1}\kappa^3\frac{s\log(p)}{n}\\ &\lesssim \mathcal{E}_4\lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Further expanding the term J_1:

What remains is to bound the first term of (6); i.e. the term J_1. First, note that by Assumption 3, each vector $X_i \in \mathbb{R}^p$ is of the form $X_i = \Sigma^{1/2} Y_i$, where $\mathbb{E} Y_i Y_i^\top = I$. Let X be the $n \times p$ matrix whose rows are the X_i's; it follows that $X = Y \Sigma^{1/2}$. Let Y be partitioned via $Y = [Y_J, Y_{J^c}]$, where Y_J is the $n \times s$ matrix of variables corresponding to those in J, and Y_{J^c} is the $n \times p - s$ matrix of the other variables. Define through slight abuse of notation the matrix $\Sigma^{1/2}_{JJ^c} := (\Sigma^{1/2})_{JJ^c}$. With these notations in place, we observe that since $\hat{\Sigma} = \frac{1}{n}(X^\top X)$ we have the block structure

$$\hat{\Sigma}_{JJ} = \frac{1}{n}\Big((\Sigma^{1/2})_{JJ} Y_J^\top Y_J (\Sigma^{1/2})_{JJ} + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} + (\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top \Big).$$

Therefore, we observe that

$$\begin{aligned}(\hat{\Sigma}_{JJ} - \Sigma_{JJ})\tilde{U}_J = \frac{1}{n}\Big(&(\Sigma^{1/2})_{JJ} (Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ} + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} \\ &+ (\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top + \Sigma^{1/2}_{JJ^c} (Y_{J^c}^\top Y_{J^c} - nI)(\Sigma^{1/2}_{JJ^c})^\top \Big)\tilde{U}_J. \qquad (7)\end{aligned}$$

Here the identity matrices are of size s and $p - s$ respectively in order of appearance. In light of the structure in (7), define the matrices

$$\begin{aligned} K_1 &:= \frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\tilde{U}_J; \\ K_2 &:= \frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ}\tilde{U}_J; \\ K_3 &:= \frac{1}{n}(\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top \tilde{U}_J; \\ K_4 &:= \frac{1}{n}\Sigma^{1/2}_{JJ^c}(Y_{J^c}^\top Y_{J^c} - nI)(\Sigma^{1/2}_{JJ^c})^\top \tilde{U}_J. \end{aligned}$$

Then

$$J_1 = (\hat{\Sigma}_{JJ} - \Sigma_{JJ})\tilde{U}_J = K_1 + K_2 + K_3 + K_4.$$

We have lemmas that bound the 2 → ∞ norms of each of these matrices. Each of these bounds follows essentially the same set of steps:

1. Bound the 2 → ∞ norm using properties of the 2 → ∞ norm in terms of the maximum entry.
2. Write each entry as a sum of mean-zero subexponential random variables and use either Bernstein's inequality or the Hanson-Wright inequality (see Appendix C) to bound the result.

The proofs for these lemmas are in Sections B.5 and B.6. Recall that we define $\mathcal{E}_5$ via

$$\begin{aligned}\mathcal{E}_5 &:= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \kappa\sqrt{\frac{k\log(p)}{n}} \\ &\equiv \frac{1}{\lambda_k}\left(\frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \lambda_1\sqrt{\frac{k\log(p)}{n}}\right).\end{aligned}$$

Lemma 7 (The matrix K_1). *The matrix K_1 satisfies*

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\tilde{U}\|_{2\to\infty} &\lesssim \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \lambda_1\sqrt{\frac{k\log(p)}{n}} \\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Lemma 8 (The matrix K_2). *The matrix K_2 satisfies*

$$\begin{aligned}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}\tilde{U}\|_{2\to\infty} &\lesssim \lambda_1\sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\frac{s\log(p)}{n}\\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1-O(p^{-4})$.

Lemma 9 (The matrices K_3 and K_4). *The matrices K_3 and K_4 satisfy*

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}Y_J^\top Y_{J^c}(\Sigma^{1/2}_{JJ^c})^\top\tilde{U}_J\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\\ &\lesssim \mathcal{E}_5\lambda_k;\\ \|\frac{1}{n}\Sigma^{1/2}_{JJ^c}(Y_{J^c}^\top Y_{J^c}-nI)(\Sigma^{1/2}_{JJ^c})^\top\tilde{U}\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1-O(p^{-3})$.

Putting it together:

We are now ready to complete the proof. We have that

$$\begin{aligned}\|\tilde{U}_J-U_JW_*\|_{2\to\infty} &\leq \left\|\left(\tilde{U}_J\tilde{\Lambda}-U_J\Lambda U^\top\hat{U}\right)\tilde{\Lambda}^{-1}\right\|_{2\to\infty} + \|T_1\|_{2\to\infty}+\|T_2\|_{2\to\infty}\\ &\leq \left\|\left(\tilde{U}_J\tilde{\Lambda}-U_J\Lambda U^\top\hat{U}\right)\tilde{\Lambda}^{-1}\right\|_{2\to\infty} + \mathcal{E}_1+\mathcal{E}_2\\ &\leq \|T_3\|_{2\to\infty}+\|T_4\|_{2\to\infty}+\mathcal{E}_1+\mathcal{E}_2\\ &\leq \|T_4\|_{2\to\infty}+\mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3\\ &\lesssim \frac{\|J_1\|_{2\to\infty}+\|J_2\|_{2\to\infty}}{\lambda_k}+\mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3\\ &\lesssim \frac{\|J_1\|_{2\to\infty}}{\lambda_k}+\mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3+\mathcal{E}_4\\ &\lesssim \frac{1}{\lambda_k}\bigg(\|K_1\|_{2\to\infty}+\|K_2\|_{2\to\infty}+\|K_3\|_{2\to\infty}+\|K_4\|_{2\to\infty}\bigg)+\mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3+\mathcal{E}_4\\ &\leq \mathcal{E}_1+\mathcal{E}_2+\mathcal{E}_3+\mathcal{E}_4+\mathcal{E}_5\end{aligned}$$

with probability at least $1-O(p^{-3})$. Consequently, by the union bound and Assumption 2, this bound holds with probability at least $1-\delta-p^{-2}$ as desired.

A.3 Proof of Theorem 1

In this section we show how Theorem 1 can be deduced from Theorem 2. We simply bound $\mathcal{E}_1$ through $\mathcal{E}_5$ using the additional assumptions introduced in Assumption 5.

Note that under Assumption 5, we have that $\lambda_{k+1}\leq\frac{\lambda}{2}$ and $\lambda_k\geq\lambda$, implying that $\lambda_k-\lambda_{k+1}\geq\frac{\lambda}{2}$. In addition $\lambda_1\leq\kappa\lambda$. Therefore,

$$\begin{aligned}\frac{\lambda_1}{\lambda_k} &\leq \frac{\kappa\lambda}{\lambda}\leq\kappa;\\ \frac{\lambda_1}{\lambda_k-\lambda_{k+1}} &\lesssim \frac{\kappa\lambda}{\lambda}\leq\kappa.\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{E}_1 &= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}} \frac{s\log(p)}{n} \|U\|_{2\to\infty} + \kappa k \sqrt{\frac{\log(p)}{n}} \|U\|_{2\to\infty} \\
&\lesssim \kappa^2 \frac{s\log(p)}{n} \|U\|_{2\to\infty} + \kappa k \sqrt{\frac{\log(p)}{n}} \|U\|_{2\to\infty} \\
&\lesssim \kappa^2 \frac{\sqrt{sk}\log(p)}{n} + \kappa \frac{k^{3/2}}{\sqrt{s}} \sqrt{\frac{\log(p)}{n}} \\
&\lesssim \kappa^2 \frac{s\log(p)}{n} + \kappa \sqrt{\frac{k\log(p)}{n}}, \qquad (8)
\end{aligned}$$

where the penultimate inequality comes from the fact that $\|U\|_{2\to\infty} \lesssim (k/s)^{1/2}$ and that $k \lesssim \sqrt{s}$. Similarly,

$$\begin{aligned}
\mathcal{E}_2 :&= \frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s\log(p)}{n} \|U\|_{2\to\infty} \\
&\lesssim \kappa^2 \frac{s\log(p)}{n} \|U\|_{2\to\infty} \\
&\lesssim \kappa^2 \frac{\sqrt{sk}\log(p)}{n} \\
&\lesssim \kappa^2 \frac{s\log(p)}{n}. \qquad (9)
\end{aligned}$$

For $\mathcal{E}_3$,

$$\begin{aligned}
\mathcal{E}_3 &= \sqrt{\frac{s\log(p)}{n}} \frac{\kappa\lambda_1^{1/2}}{\lambda_k - \lambda_{k+1}} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty} \right) \\
&\lesssim \sqrt{\frac{s\log(p)}{n}} \frac{\kappa\lambda_1^{1/2}}{\lambda_k - \lambda_{k+1}} \sqrt{\lambda_1}\|U\|_{2\to\infty} \\
&\lesssim \kappa^2 \sqrt{\frac{s\log(p)}{n}} \|U\|_{2\to\infty} \\
&\lesssim \kappa^2 \sqrt{\frac{k\log(p)}{n}} \qquad (10)
\end{aligned}$$

since $\|U\|_{2\to\infty} \lesssim (k/s)^{1/2}$. For $\mathcal{E}_4$, we have that since $\lambda_{k+1} < \lambda_k$, then

$$\begin{aligned}
\mathcal{E}_4 &= \kappa^2 \frac{\lambda_{k+1}}{\lambda_k} \sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_{k+1}}{\lambda_k} \kappa^3 \frac{s\log(p)}{n} \\
&\lesssim \kappa^2 \sqrt{\frac{k\log(p)}{n}} + \kappa^3 \frac{s\log(p)}{n}. \qquad (11)
\end{aligned}$$

Finally, for $\mathcal{E}_5$, we see that

$$\begin{aligned}
\mathcal{E}_5 :&= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}} \frac{s\log(p)}{n} + \kappa \sqrt{\frac{k\log(p)}{n}}. \\
&\lesssim \kappa^2 \frac{s\log(p)}{n} + \kappa \sqrt{\frac{k\log(p)}{n}}. \qquad (12)
\end{aligned}$$

The condition number is always larger than 1. Hence, combining (8),(9),(10),(11) and (12) completes the proof.

B Proofs of Intermediate Lemmas

In this section we collect the proofs of the Lemmas needed en route to the proof of Theorem 2. All the lemmas are self-contained and repeated for convenience.

B.1 Proofs of Lemmas 1 and 2

First, we recall the statement of Lemma 1.

Lemma 1 (Principal Submatrix Concentration)**.** *Let J be an index set of $\{1, ..., p\}$ of size s. Then*

$$\|\hat{\Sigma}_{JJ} - \Sigma_{JJ}\| \lesssim \lambda_1 \bigg(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \bigg)$$

with probability at least $1 - O(p^{-4})$.

Proof of Lemma 1. The result is similar to Amini and Wainwright (2009), but for general subgaussian ensembles as opposed to Gaussian ensembles. The proof is standard via an ε-net; we follow similarly to the proof of Theorem 6.5 in Wainwright (2019).

Let $\Delta = \hat{\Sigma}_{JJ} - \Sigma_{JJ}$. First take an 1/8-net of the S^{s-1} sphere; denote these vectors $v_1, ..., v_N$, where $N \leq 17^s$ (see Example 5.8 in Wainwright (2019)). Then for any s-unit vector v, there exists some vector v_j of distance at most $\varepsilon = \frac{1}{8}$ to v. Therefore

$$\langle v, \Delta v \rangle = \langle v_j, \Delta v_j \rangle + 2\langle (v - v_j), \Delta v_j \rangle + \langle v - v_j, \Delta(v - v_j) \rangle.$$

Hence, we see that by the triangle inequality and Cauchy-Schwarz,

$$\begin{aligned} |\langle v, \Delta v \rangle| &\leq |\langle v_j, \Delta v_j \rangle| + 2\|\Delta\| \|v - v_j\| \|v_j\| + \|\Delta\| \|v - v_j\|^2 \\ &\leq |\langle v_j, \Delta v_j \rangle| + \frac{1}{2}\|\Delta\|, \end{aligned}$$

where the final inequality comes from the fact that v_j is at most distance $\frac{1}{8}$ to v. Letting v denote the unit vector achieving $\sup \langle v, Qv \rangle$ and rearranging we have that

$$\|\Delta\| \leq 2|\langle v_j, \Delta v_j \rangle| \leq 2 \max_{1 \leq i \leq n} |\langle v_i, \Delta v_i \rangle|.$$

So we therefore have that

$$\mathbb{E}(\exp(\lambda\|\Delta\|)) \leq \mathbb{E}\bigg(\exp(2\lambda \max_{1 \leq i \leq N} |\langle v_i, \Delta v_i \rangle|) \bigg) \leq \sum_{i=1}^{N} \bigg(\mathbb{E}(\exp(2\lambda \langle v_i, \Delta v_i \rangle)) + \mathbb{E}(\exp(-2\lambda \langle v_i, \Delta v_i \rangle)) \bigg).$$

We now bound the mgf above, which is the primary technical difference between this and Theorem 6.5 in Wainwright (2019). Denote $X_i[J]$ as the vector X_i with only the components in J, and let u be an arbitrary unit vector. From the assumption the X_i's are iid we have that

$$\begin{aligned} \mathbb{E} \exp(tu^\top \Delta u) &= \prod_{i=1}^{n} \mathbb{E}_{X_i} \left[\exp \left(\frac{t}{n} [(X_i[J]^\top u)^2 - u^\top \Sigma_{JJ} u] \right) \right] \\ &= \left(\mathbb{E}_{X_1} \left[\exp \left(\frac{t}{n} [(X_1[J]^\top u)^2 - u^\top \Sigma_{JJ} u] \right) \right] \right)^n. \end{aligned}$$

Let ε be a Rademacher random variable independent of X_1. Then by the contraction property of Rademacher random variables,

$$\begin{aligned} \mathbb{E}_{X_1} \left[\exp \left(\frac{t}{n} [(X_1[J]^\top u)^2 - u^\top \Sigma_{JJ} u] \right) \right] &\leq \mathbb{E}_{X_1, \varepsilon} \left[\exp \left(\frac{2t}{n} \varepsilon ((X_1[J])^\top u)^2 \right) \right] \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{2t}{n} \right)^k \mathbb{E}(\varepsilon^k (X_1[J]^\top u)^{2k}) \\ &= 1 + \sum_{k=1}^{\infty} \frac{1}{(2k)!} \left(\frac{2t}{n} \right)^{2k} \mathbb{E}((X_1[J]^\top u)^{4k}) \end{aligned}$$

where the first is by the series expansion for the exponential, and the second is by noting that the ε are rademacher and hence have vanishing odd moments.

Note that by assumption the X_i's can be written as $X_i = \Sigma^{1/2} Y_i$ for some independent Y_i's satisfying $\|Y_{ij}\|_{\psi_2} \leq 1$. Then $\|\Sigma^{1/2} Y_i\|_{\psi_2} \leq \sqrt{\lambda_1}$. Hence, by equivalence of the subgaussian norm, the moments satisfy

$$\mathbb{E}((X_1[J]^\top u)^{4k}) \leq \frac{(4k)!}{2^{2k}(2k)!}(\sqrt{8}e\lambda_1^{1/2})^{4k}.$$

From this, we deduce

$$\begin{aligned} 1 + \sum_{k=1}^{\infty} \frac{1}{(2k)!}\big(\frac{2t}{n}\big)^{2k} \mathbb{E}((X_1[J]^\top u)^{4k}) &\leq 1 + \sum_{k=1}^{\infty} \frac{1}{(2k)!}\left(\frac{2t}{n}\right)^{2k} \frac{(4k)!}{2^{2k}(2k)!}(\sqrt{8}e\lambda_1^{1/2})^{4k} \\ &\leq 1 + \sum_{k=1}^{\infty}\left(\frac{16t}{n}e^2\lambda_1\right)^{2k} \end{aligned}$$

which is a geometric series. Hence, since $\frac{1}{1-a} \leq e^{2a}$ for all $a \in [0, 1/2]$, we have that

$$1 + \sum_{k=1}^{\infty}\left(\frac{16t}{n}e^2\lambda_1\right)^{2l} \leq \exp\left(2\big[\frac{16t}{n}e^2\lambda_1\big]^2\right)$$

for all $|t| < \frac{n}{32e^2\lambda_1}$. Therefore, we have shown

$$\mathbb{E}\exp(tu^\top \Delta u) \leq \exp\left(512\frac{t^2}{n}e^4\lambda_1^2\right).$$

From here, using the sum, we have that for all $|t| < \frac{n}{64e^2\lambda_1}$ that

$$\begin{aligned} \mathbb{E}(\exp(t\|\Delta\|)) &\leq \sum_{i=1}^{N}\Big(\mathbb{E}(\exp(2t\langle v_i, \Delta v_i\rangle)) + \mathbb{E}(\exp(-2t\langle v_i, \Delta v_i\rangle))\Big) \\ &\leq 2Ne^{2048\frac{t^2}{n}e^4\lambda_1^2} \\ &\leq \exp(C\frac{t^2\lambda_1^2}{n} + 4s), \end{aligned}$$

since $2(17^s) \leq e^{4s}$. Therefore, by the Chernoff bound,

$$\mathbb{P}\Big(\|\Delta\| > \eta\Big) \leq \exp\left(C\frac{t^2\lambda_1^2}{n} + 4s - \eta t\right).$$

Minimizing over t shows that

$$t = \frac{n\eta}{2C\lambda_1^2}$$

is the minimizer provided that $\eta < \frac{C\lambda_1}{32e^2}$. Plugging this value of t in yields

$$\begin{aligned} \mathbb{P}\Big(\|\Delta\| > \eta\Big) &\leq \exp\left(4s - \frac{\eta^2 n}{4C\lambda_1^2}\right) \\ &= \exp\left[n\left(\frac{4s}{n} - \frac{\eta^2}{4C\lambda_1^2}\right)\right]. \end{aligned}$$

Suppose $\eta = C_2\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{4\log(p)}{n}}\right)$ for some sufficiently large constant C_2. Note that Assumption 1 ensures that this choice of η satisfies $\eta < \frac{C\lambda_1}{32e^2}$ since $s/n = o(1)$ and $\log(p)/n = o(1)$. Therefore, with this choice of η, we have that

$$\begin{aligned} \exp\left[n\left(\frac{4s}{n} - \frac{\eta^2}{4C\lambda_1^2}\right)\right] &\leq \exp(-4\log(p)) \\ &\leq p^{-4}. \end{aligned}$$

Consequently, recalling that $\Delta = \hat{\Sigma}_{JJ} - \Sigma_{JJ}$ we have that

$$\mathbb{P}\bigg[\|\hat{\Sigma}_{JJ} - \Sigma_{JJ}\| > C_2 \lambda_1 \bigg(\sqrt{\frac{s}{n}} + \sqrt{\frac{4\log(p)}{n}} \bigg) \bigg] \leq p^{-4}$$

as desired. □

Again, we recall the statement of Lemma 2.

Lemma 2 (Existence of an Eigengap)**.** *Under the event in Lemma 1 and Assumption 4, the eigenvalues of* $\hat{\Sigma}_{JJ}$ *and* Σ_{JJ} *satisfy*

$$\lambda_k - \tilde{\lambda}_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8}; \qquad \tilde{\lambda}_k - \lambda_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8};$$
$$\tilde{\lambda}_k \geq \frac{\lambda_k}{4}.$$

Consequently, this bound holds with probability at least $1 - O(p^{-4})$.

Proof of Lemma 2. By Weyl's inequality, the event in Lemma 1 implies that for all $1 \leq i \leq s$ that

$$|\lambda_i - \tilde{\lambda}_i| \leq C\lambda_1 \bigg(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \bigg).$$

Note that Σ_{JJ} is a principal submatrix of Σ; hence its eigenvalues satisfy $\lambda_i(\Sigma_{JJ}) \leq \lambda_i$ for all $i \geq k+1$ (when $1 \leq i \leq k$ we have equality). Therefore, By Assumption 4, we have that

$$\begin{aligned}
\lambda_k - \tilde{\lambda}_{k+1} &\geq \lambda_k - \lambda_{k+1}(\Sigma_{JJ}) - C\lambda_1 \bigg(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \bigg) \\
&\geq \lambda_k - \lambda_{k+1} - C\lambda_1 \bigg(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \bigg) \\
&\geq \frac{7}{8}(\lambda_k - \lambda_{k+1}) \\
&\geq \frac{\lambda_k - \lambda_{k+1}}{8},
\end{aligned}$$

and similarly for $\tilde{\lambda}_k - \lambda_{k+1}$. For the final bound,

$$\begin{aligned}
\tilde{\lambda}_k &\geq \lambda_k - C\lambda_1 \bigg(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \bigg) \\
&\geq \lambda_k - \lambda_k/8 \\
&\geq \frac{\lambda_k}{4},
\end{aligned}$$

which completes the proof. □

B.2 Proof of Lemmas 3 and 4

First we will recall the statement of Lemma 3.

Lemma 3 (Bound on T_1)**.** *We have that*

$$\begin{aligned}
\|U_J(\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J \tilde{\Lambda})\tilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})} \frac{s\log(p)}{n} \\
&\quad + \frac{k\lambda_1 \|U\|_{2\to\infty}}{\lambda_k} \sqrt{\frac{\log(p)}{n}} \\
&\equiv \mathcal{E}_1
\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Proof of Lemma 3. Note that by properties of the $2 \to \infty$ norm, we have

$$\|U_J(\Lambda U_J^\top \tilde{U} - U_J^\top \tilde{U}_J \tilde{\Lambda})\tilde{\Lambda}^{-1}\|_{2\to\infty} \leq \|U_J\|_{2\to\infty}\|\Lambda U_J^\top \tilde{U} - U_J^\top \tilde{U}_J\tilde{\Lambda}\|\hat{\lambda}_k^{-1}. \tag{13}$$

We note that $\lambda_k \lesssim \tilde{\lambda}_k$ with probability $1 - O(p^{-4})$ by Lemma 2. Furthermore, by the eigenvector equation,

$$\begin{aligned}\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J \tilde{\Lambda} &= (U_J\Lambda)^\top \tilde{U}_J - U_J^\top \tilde{U}_J\tilde{\Lambda} \\ &= (\Sigma_{JJ}U_J)^\top \tilde{U} - U_J^\top \hat{\Sigma}_{JJ}\tilde{U}_J \\ &= U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})\tilde{U}_J.\end{aligned}$$

In addition,

$$U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})\tilde{U}_J = U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J U_J^\top \tilde{U}_J + U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})(I - U_J U_J^\top)\tilde{U}_J.$$

The second term satisfies

$$\|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})(I - U_J U_J^\top)\tilde{U}_J\| \leq \|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})\|\|(I - U_J U_J^\top)\tilde{U}_J\|.$$

Note that

$$\begin{aligned}\|(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J\| &\leq \|\Sigma_{JJ} - \hat{\Sigma}_{JJ}\|\|U_J\| \\ &\leq \|\Sigma_{JJ} - \hat{\Sigma}_{JJ}\|\end{aligned}$$

since U_J has orthonormal columns. Therefore, by Lemma 1,

$$\|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})\| \lesssim \lambda_1\left(\sqrt{\frac{s\log(p)}{n}}\right). \tag{14}$$

Note that $\|(I - U_J U_J^\top)\tilde{U}_J\| \lesssim \|\sin\Theta(U_J, \tilde{U}_J)\| \lesssim \|U_J U_J^\top - \tilde{U}_J\tilde{U}_J^\top\|$ (see Lemma 10 in Appendix C). Therefore, by Proposition 1, we have that

$$\|(I - U_J U_J^\top)\tilde{U}_J\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}\left(\sqrt{\frac{s\log(p)}{n}}\right). \tag{15}$$

In summary, we have shown so far that by (13), (14), and (15),

$$\begin{aligned}\|U_J(\Lambda U_J^\top \tilde{U}_J - U_J^\top \tilde{U}_J\tilde{\Lambda})\tilde{\Lambda}^{-1}\|_{2\to\infty} \lesssim\ & \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})}\frac{s\log(p)}{n} \\ &+ \frac{\|U\|_{2\to\infty}}{\lambda_k}\|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J U_J^\top \tilde{U}_J\|.\end{aligned}$$

Therefore, we focus on bounding the term

$$\|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J U_J^\top \tilde{U}_J\|.$$

Naively, $\|U_J^\top \tilde{U}_J\| \leq 1$ so by submultiplicativity, we have that

$$\|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J U_J^\top \tilde{U}_J\| \leq \|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J\|.$$

For any indices i and k, the entry of the above matrix can be written as

$$\frac{1}{n}\sum_{l=1}^n \langle (U_J)_{\cdot i}, (X_l X_l^\top - \mathbb{E}(X_l X_l^\top))(U_J)_{\cdot k}\rangle = \frac{1}{n}\sum_{l=1}^n\Big(((U_J)_{\cdot i}^\top X_l)(X_l^\top(U_J)_{\cdot k}) - (U_J)_{\cdot i}^\top\Sigma(U_J)_{\cdot k}\Big).$$

This is a sum of independent, mean-zero subexponential random variables. Therefore, to apply the generalized Bernstein inequality (see Theorem 3 in Appendix C), we need to find the ψ_1 norm of the above random variable.

By properties of the ψ_1 norm, we have that

$$\begin{aligned}\|((U_J)_{\cdot i}^\top X_l)(X_l^\top (U_J)_{\cdot k}) - (U_J)_{\cdot j}^\top \Sigma (U_J)_{\cdot i}\|_{\psi_1} &\leq C\|((U_J)_{\cdot i}^\top X_l)(X_l^\top (U_J)_{\cdot k})\|_{\psi_1} \\ &\leq C\|(U_J)_{\cdot i}^\top X_l\|_{\psi_2}\|X_l^\top (U_J)_{\cdot k}\|_{\psi_2} \\ &= C\|(U_J)_{\cdot i}^\top \Sigma^{1/2} Y_l\|_{\psi_2}\|Y_l^\top \Sigma^{1/2}(U_J)_{\cdot k}\|_{\psi_2} \\ &= C\sqrt{\lambda_i\lambda_k}\|(U_J)_{\cdot i}^\top Y_l\|_{\psi_2}\|(U_J)_{\cdot k}^\top Y_l\|_{\psi_2} \\ &\leq C\sqrt{\lambda_j\lambda_k} \\ &\leq C\lambda_1\end{aligned}$$

since $X_l = \Sigma^{1/2} Y_l$, the $(U_J)_{\cdot i}$ are the eigenvectors of Σ and the vectors Y are assumed to be subgaussian with unit ψ_2 norm. Therefore, by the generalized Bernstein inequality (Theorem 3), we have that for fixed i, k, that

$$\mathbb{P}\bigg(|\frac{1}{n}\sum_{l=1}^n \langle (U_J)_{\cdot i}, (X_l X_l^\top - \mathbb{E}(X_l X_l^\top))(U_J)_{\cdot k}|\rangle \geq t\bigg) \leq 2\exp\bigg[-c_0 n \min\bigg(\frac{t^2}{(\lambda_1)^2}, \frac{t}{\lambda_1}\bigg)\bigg].$$

Since $\log(k) \ll \log(p)$, taking $t = C\lambda_1\sqrt{\frac{2\log(k)+4\log(p)}{n}}$ for some constant C yields that

$$\begin{aligned}|(U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J)_{ik}| &\leq C\lambda_1\sqrt{\frac{2\log(k)+4\log(p)}{n}} \\ &\lesssim \lambda_1\sqrt{\frac{\log(p)}{n}}\end{aligned}$$

with probability at least $1 - O(p^{-4}k^{-2})$. Therefore,

$$\begin{aligned}\|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J\| &\leq \|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J\|_F \\ &\leq k\|U_J^\top(\Sigma_{JJ} - \hat{\Sigma}_{JJ})U_J\|_{\max} \\ &\leq Ck\lambda_1\sqrt{\frac{2\log(k)+4\log(p)}{n}} \\ &\lesssim k\lambda_1\sqrt{\frac{\log(p)}{n}}\end{aligned}$$

with probability at least $1-O(p^{-4})$ by taking a union bound over all k^2 entries. Therefore, putting it all together, we see that

$$\begin{aligned}\|U_J(\Lambda U_J^\top \tilde{U} - U_J^\top \tilde{U}_J\tilde{\Lambda})\tilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})}\frac{s\log(p)}{n} \\ &\quad + \frac{k\lambda_1\|U\|_{2\to\infty}}{\lambda_k}\sqrt{\frac{\log(p)}{n}}\end{aligned}$$

with probability at least $1 - O(p^{-4})$ as desired. □

Now we prove Lemma 4.

Lemma 4 (Bound on T_2)**.** *We have that*

$$\begin{aligned}\|U_J(W_* - U_J^\top \tilde{U}_J)\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2}\frac{s\log(p)}{n} \\ &\equiv \mathcal{E}_2\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Proof of Lemma 4. This proof follows similarly to ideas in Cape et al. (2019a); Abbe et al. (2020b); Lei (2019).

By properties of the $2 \to \infty$ norm, we have

$$\|U_J(W_* - U_J^\top \tilde{U}_J)\|_{2\to\infty} \leq \|U_J\|_{2\to\infty}\|W_* - U_J^\top \tilde{U}_J\|.$$

We will now analyze the term inside the spectral norm. Note that W_* is the Frobenius-optimal Procrustes transformation. Let $V_1 \Sigma V_2^\top$ be the SVD of $U_J^\top \tilde{U}_J$. Then Σ contains the sines of the canonical angles between U_J and $\tilde{U}_J$ (see Bhatia (1997) or Stewart and Sun (1990) for details; Lemma 10 in Appendix C also contains equivalent expressions for the $\sin\Theta$ distances). Then, letting θ_j be the canonical angles and $\sigma_j = \cos(\theta_j)$,

$$\begin{aligned}
\|W_* - U_J^\top \tilde{U}_J\| &= \|V_1 V_2^\top - V_1 \Sigma V_2^\top\| \\
&= \|I - \Sigma\| \\
&= \max_{1\leq j \leq k} |(1-\sigma_j)| \\
&\leq \max_{1\leq j \leq k} (1-\sigma_j^2) \\
&= \max_j \sin^2(\theta_j) \\
&= \|U_J U_J^\top - \tilde{U}_J \tilde{U}_J^\top\|^2 \\
&\lesssim \frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s\log(p)}{n}.
\end{aligned}$$

with probability at least $1 - O(p^{-4})$ by Proposition 1. □

B.3 Proof of Lemma 5

Recall the statement of Lemma 5.

Lemma 5 (Bound on T_3)**.** *We have that*

$$\begin{aligned}
\|U_J U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \sqrt{\frac{s\log(p)}{n}} \frac{\lambda_1^{3/2}}{\lambda_k(\lambda_k - \lambda_{k+1})} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty}\right) \\
&\equiv \mathcal{E}_3
\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Proof of Lemma 5. Note that since $U_J^\top \Sigma_{JJ} = \Lambda U_J^\top$, we have that

$$\begin{aligned}
\|U_J U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1}\|_{2\to\infty} &\leq \frac{\|U\|_{2\to\infty}}{\tilde{\lambda}_k}\|U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\| \\
&\leq \frac{\|U\|_{2\to\infty}}{\tilde{\lambda}_k}\|\Lambda U_J^\top(\tilde{U}_J \tilde{U}_J^\top - U_J U_J^\top)\|.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|U_J U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1}\|_{2\to\infty} &\leq \frac{\|U\Lambda^{1/2}\|_{2\to\infty}}{\tilde{\lambda}_k}\|\Lambda^{1/2} U_J^\top(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\| \\
&\leq \frac{\sqrt{\|\Sigma\|_{\max}}}{\tilde{\lambda}_k}\|\Lambda^{1/2} U_J^\top(\tilde{U}_J \tilde{U}_J^\top - U_J U_J^\top)\|,
\end{aligned}$$

where the term $\|\Sigma\|_{\max}$ comes from the fact that

$$|U_J U_J^\top \Sigma_{i,j}| = |\langle (U\Lambda^{1/2})_i, (U\Lambda^{1/2})_j\rangle|,$$

and hence that

$$\begin{aligned}
\|U|\Lambda|^{1/2}\|_{2\to\infty} &= \max_i \sqrt{\langle (U\Lambda^{1/2})_i, (U\Lambda^{1/2})_i\rangle} \\
&\leq \max_i \sqrt{|(U_J U_J^\top \Sigma)_{ii}|} \\
&\leq \max_{i,j} \sqrt{|\Sigma_{ij}|},
\end{aligned}$$

since the eigenvalues of Σ are all positive. Therefore,

$$\|U_J U_J^\top \Sigma_{JJ}(\tilde{U}_J - U_J U_J^\top \tilde{U}_J)\tilde{\Lambda}^{-1}\|_{2\to\infty} \leq \frac{1}{\tilde{\lambda}_k} \min\bigg(\sqrt{\lambda_1}\|U\|_{2\to\infty}\|\Lambda^{1/2}U_J^\top(\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top)\|, \\ \|\Sigma\|_{\max}^{1/2}\|\Lambda^{1/2}U_J^\top(\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top)\|\bigg) \tag{16}$$

Therefore, what remains is to analyze

$$\|\Lambda^{1/2}U_J^\top(\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top)\|.$$

To find this bound, we will represent the difference $\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top$ using the holomorphic functional calculus as done in Lei (2019) for the spiked Wigner matrix setting. This technique has been used extensively in studying eigenvector perturbation; e.g. Mao et al. (2020); Lei (2019); Koltchinskii and Xia (2016); Xia (2021); Wahl (2019a,b). More specifically, let $\mathcal{C}$ denote a contour on the complex plane with real part ranging from $\lambda_k - \eta$ to $\lambda_1 + \eta$, and with imaginary part ranging from $-\gamma$ to γ. Then, for a proper choice of η, the top k eigenvalues of both Σ_{JJ} and $\hat{\Sigma}_{JJ}$ lie in $\mathcal{C}$, and one can write the difference of the spectral projections via a complex integral

$$\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top = -\left[\frac{1}{2\pi i}\oint_{\mathcal{C}}(\hat{\Sigma}_{JJ} - zI)^{-1}dz - \frac{1}{2\pi i}\oint_{\mathcal{C}}(\Sigma_{JJ} - zI)^{-1}dz\right]$$

by the residue theorem (e.g. (Greene and Krantz, 2006)). Using the identity $A^{-1} - B^{-1} = B^{-1}(A - B)A^{-1}$, and assuming the real number η is chosen appropriately so that the contours are the same, the integrals can be combined to arrive at the expression

$$\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top = -\frac{1}{2\pi i}\oint_{\mathcal{C}}(\Sigma_{JJ} - zI)^{-1}(\hat{\Sigma}_{JJ} - \Sigma_{JJ})(\hat{\Sigma}_{JJ} - zI)^{-1}dz.$$

Premultiplying by $\Lambda^{1/2}U_J^\top$ yields (formally) that

$$\|\Lambda^{1/2}U_J^\top(\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top)\| = \frac{1}{2\pi}\left\|\oint_{\mathcal{C}}\Lambda^{1/2}U_J^\top(\Sigma_{JJ} - zI)^{-1}(\hat{\Sigma}_{JJ} - \Sigma_{JJ})(\hat{\Sigma}_{JJ} - zI)^{-1}dz\right\|.$$

Note that the matrix is diagonalizable by the same eigenvectors as Σ_{JJ}, so that

$$\begin{aligned} U_J^\top(\Sigma_{JJ} - zI)^{-1} &= U_J^\top(U_J(\Lambda - zI)^{-1}U_J^\top) + U_J^\top(U_\perp(\Lambda_\perp - zI)^{-1}U_\perp^\top \\ &= (\Lambda - zI)^{-1}U_J^\top \end{aligned}$$

by orthonormality, where $U_\perp$ are defined as the $s \times s$ completion of U_J such that $[U_J, U_\perp]$ is an $s \times s$ orthogonal matrix. Therefore, we have

$$\|\Lambda^{1/2}U_J^\top(\tilde{U}_J\tilde{U}_J^\top - U_J U_J^\top)\| = \frac{1}{2\pi}\left\|\oint_{\mathcal{C}}\Lambda^{1/2}(\Lambda - zI)^{-1}U_J^\top(\hat{\Sigma}_{JJ} - \Sigma_{JJ})[\tilde{U}, \tilde{U}_\perp](\hat{\Lambda}_{all} - zI)^{-1}dz\right\|,$$

where $\hat{\Lambda}_{all}$ is the diagonal matrix of all the eigenvalues of $\hat{\Sigma}_{JJ}$.

The rest of the proof mirrors closely that of Lemma A.8 in Lei (2019). Recall that in order to do all these manipulations, we required that the parameter η was chosen such that the contour $\mathcal{C}$ contains the top k eigenvalues of $\hat{\Sigma}_{JJ}$ and Σ_{JJ}. In fact, Lemma 2 shows that the choice

$$\eta := \frac{\lambda_k - \lambda_{k+1}}{4}$$

suffices. To see this, note that by Lemmas 1 and 2,

$$\begin{aligned} |\tilde{\lambda}_k - \lambda_k| &\leq \frac{\lambda_k - \lambda_{k+1}}{8}; \\ |\tilde{\lambda}_{k+1} - \lambda_{k+1}| &\leq \frac{\lambda_k - \lambda_{k+1}}{8}, \end{aligned}$$

so that the interval $\lambda_k \pm \eta$ contains $\tilde{\lambda}_k$, the interval $\lambda_k \pm \eta$ does not contain $\tilde{\lambda}_{k+1}$, and both $\tilde{\lambda}_k$ and $\tilde{\lambda}_{k+1}$ satisfy

$$\begin{aligned} |\lambda_k - \tilde{\lambda}_k - \eta| &\geq \eta/2 \\ |\lambda_k - \tilde{\lambda}_{k+1} - \eta| &\geq \eta/2. \end{aligned}$$

Therefore, the top k eigenvalues of $\hat{\Sigma}_{JJ}$ lie within $\mathcal{C}$ with high probability and the bottom eigenvalues lie outside of it. With this particular choice of η, we can proceed to bound the integrand along the contour $\mathcal{C}$. We will conduct the rest of the analysis assuming that this event holds; it does with probability at least $1 - O(p^{-4})$.

Define $a := \lambda_k - \eta$ and $b := \lambda_1 + \eta$. We decompose the contour $\mathcal{C}$ into the following sets

$$\begin{aligned} \mathcal{C}_1 &:= \{z = a + xi, x \in (-\gamma, \gamma)\} \qquad & \mathcal{C}_2 &:= \{z = x + \gamma i : x \in [a, b]\} \\ \mathcal{C}_3 &:= \{z = b + xi, x \in (-\gamma, \gamma)\} \qquad & \mathcal{C}_4 &:= \{z = x - \gamma i : x \in [a, b]\}. \end{aligned}$$

Let $\mathcal{I}(z)$ be the integrand. Observe that

$$\left\| \oint_{\mathcal{C}} \mathcal{I}(z) dz \right\| \leq \oint_{\mathcal{C}_1} \left\| \mathcal{I}(z) dz \right\| + \oint_{\mathcal{C}_2} \left\| \mathcal{I}(z) dz \right\| + \oint_{\mathcal{C}_4} \left\| \mathcal{I}(z) dz \right\| + \oint_{\mathcal{C}_4} \left\| \mathcal{I}(z) dz \right\|.$$

Therefore, we bound the above integrals directly. The tricky analysis will be along $\mathcal{C}_1$ and $\mathcal{C}_3$; we will show that the integral along $\mathcal{C}_2$ and $\mathcal{C}_4$ tend to zero for large γ. To this end, we will focus on $\mathcal{C}_1$ first. Note that

$$\begin{aligned} &\oint_{\mathcal{C}_1} \left\| \Lambda^{1/2} (\Lambda - zI)^{-1} U_J^\top (\hat{\Sigma}_{JJ} - \Sigma_{JJ}) [\tilde{U}, \tilde{U}_\perp] (\hat{\Lambda}_{all} - zI)^{-1} \right\| dz \qquad (17) \\ &\leq \oint_{\mathcal{C}_1} \left\| \Lambda^{1/2} (\Lambda - zI)^{-1} \right\| \left\| U_J^\top (\hat{\Sigma}_{JJ} - \Sigma_{JJ}) [\tilde{U}, \tilde{U}_\perp] \right\| \left\| (\hat{\Lambda}_{all} - zI)^{-1} \right\| dz \\ &\leq \left\| U_J^\top (\hat{\Sigma}_{JJ} - \Sigma_{JJ}) [\tilde{U}, \tilde{U}_\perp] \right\| \int_{-\gamma}^{\gamma} \left\| \Lambda^{1/2} (\Lambda - (a + xi) I)^{-1} \right\| \left\| (\hat{\Lambda}_{all} - (a + xi) I)^{-1} \right\| dx. \end{aligned}$$

First, recall the definition of $a := \lambda_k - \eta$. The term on the right-most side satisfies

$$\left\| (\hat{\Lambda}_{all} - (a + xi) I)^{-1} \right\| \leq \frac{1}{\sqrt{(\eta)^2/4 + x^2}}$$

for all x since $(\hat{\lambda}_i - a) \geq \eta/2$. Therefore, we are left to bound the middle term, for which we must bound

$$\max_{1 \leq i \leq k} \frac{\lambda_i^{1/2}}{\sqrt{(\lambda_i - a)^2 + x^2}}.$$

Define the function

$$g(u; x, a) := \frac{u}{\sqrt{(u - a)^2 + x^2}}.$$

Then

$$\max_{1 \leq i \leq k} \frac{\lambda_i^{1/2}}{\sqrt{(\lambda_i - a)^2 + x^2}} \leq \sup_{u \geq a + \eta} \bigg(g(u; x, a) \bigg)^{1/2} \frac{1}{(\eta^2 + x^2)^{1/4}}.$$

The details of the function g are carried out in Lei (2019); the analysis therein implies

$$\sup_{u \geq a + \eta} g(u; x, a) \leq \frac{a + \eta}{\sqrt{\eta^2 + x^2}} \mathbb{I}_{|x| \leq \sqrt{a\eta}} + \sqrt{\frac{a + \eta}{\eta}} \mathbb{I}_{|x| > \sqrt{a\eta}}.$$

Therefore the integral from (17) satisfies

$$
\begin{aligned}
&\int_{-\gamma}^{\gamma} \|\Lambda^{1/2}(\Lambda - (a+xi)I)^{-1}\| \|(\hat{\Lambda}_{all} - (a+xi)I)^{-1}\| dx \\
&\quad \leq \int_{-\gamma}^{\gamma} \frac{1}{\sqrt{\eta^2/4 + x^2}} \frac{1}{(\eta^2+x^2)^{1/4}} \left(\frac{a+\eta}{\sqrt{\eta^2+x^2}} \mathbb{I}_{|x|\leq\sqrt{a\eta}} + \sqrt{\frac{a+\eta}{\eta}} \mathbb{I}_{|x|>\sqrt{a\eta}} \right)^{1/2} dx \\
&\quad \leq \int_{|x|\leq\sqrt{a\eta}} \frac{4}{(\eta^2+x^2)^{3/4}} \left(\frac{a+\eta}{\sqrt{\eta^2+x^2}} \right)^{1/2} dx + \int_{|x|>\sqrt{a\eta}} \frac{4}{(\eta^2+x^2)^{3/4}} \left(\sqrt{\frac{a+\eta}{\eta}} \right)^{1/2} dx \\
&\quad \leq 4\sqrt{a+\eta} \int_{|x|\leq\sqrt{a\eta}} \frac{1}{\eta^2+x^2} dx + 4\left(\frac{a+\eta}{\eta}\right)^{1/4} \int_{|x|>\sqrt{a\eta}} \frac{1}{(\eta^2+x^2)^{3/4}} dx \\
&\quad \leq 8\sqrt{a+\eta} \int_{0}^{\sqrt{a\eta}} \frac{1}{\eta^2+x^2} dx + 8\left(\frac{a+\eta}{\eta}\right)^{1/4} \int_{\sqrt{a\eta}}^{\infty} \frac{1}{(\eta^2+x^2)^{3/4}} dx \\
&\quad \leq 8\frac{\sqrt{a+\eta}}{\eta} \int_{0}^{\sqrt{a/\eta}} \frac{1}{1+u^2} du + 8\left(\frac{a+\eta}{\eta}\right)^{1/4} \frac{1}{\eta^{1/2}} \int_{\sqrt{a/\eta}}^{\infty} \frac{1}{(1+u^2)^{3/4}} du \\
&\quad \leq 8\frac{\sqrt{a+\eta}}{\eta} 2\pi + 8\left(\frac{a+\eta}{\eta}\right)^{1/4} \frac{1}{\eta^{1/2}} \int_{\sqrt{a/\eta}}^{\infty} \frac{1}{u^{3/2}} du \\
&\quad \leq 16\pi \frac{\sqrt{a+\eta}}{\eta} + 8\left(\frac{a+\eta}{\eta}\right)^{1/4} \frac{1}{\eta^{1/2}} \frac{2}{(a/\eta)^{1/2}} \\
&\quad \lesssim \frac{\sqrt{a+\eta}}{\eta} + \left(\frac{a+\eta}{a}\right)^{1/4} \frac{1}{a^{1/2}}.
\end{aligned}
$$

Recall that $a + \eta = \lambda_k$; $\eta = (\lambda_k - \lambda_{k+1})/4$. With these, the bound becomes (up to constants)

$$
\begin{aligned}
&\oint_{\mathcal{C}_1} \left\| \Lambda^{1/2}(\Lambda - zI)^{-1} U_J^\top (\hat{\Sigma}_{JJ} - \Sigma_{JJ}(\tilde{U}, \tilde{U}_\perp)(\hat{\Lambda}_{all} - zI)^{-1} \right\| dz \\
&\qquad\qquad \lesssim \frac{\sqrt{\lambda_1}}{\lambda_k - \lambda_{k+1}} \|(\hat{\Sigma}_{JJ} - \Sigma_{JJ})U_J\| + \kappa^{1/4} \frac{1}{\lambda_k^{1/2}} \|(\hat{\Sigma}_{JJ} - \Sigma_{JJ})U_J\| \\
&\qquad\qquad \lesssim \frac{\sqrt{\lambda_1}}{\lambda_k - \lambda_{k+1}} \|(\hat{\Sigma}_{JJ} - \Sigma_{JJ})U_J\|.
\end{aligned}
$$

The exact same argument goes through for contour $\mathcal{C}_3$ as well. We will see that the contours along the imaginary axis tend to zero as $\gamma \to \infty$. Assuming this for the moment, by Equation (16), we see that the final bound is of the form

$$
\begin{aligned}
&\frac{1}{\lambda_k} \|\Lambda^{1/2} U_J^\top (\tilde{U}_J \tilde{U}_J^\top - U_J U_J^\top)\| \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty} \right) \\
&\qquad\qquad \lesssim \frac{\|(\hat{\Sigma}_{JJ} - \Sigma_{JJ})U_J\|}{\lambda_k} \left(\frac{\sqrt{\lambda_1}}{\lambda_k - \lambda_{k+1}} \right) \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty} \right)
\end{aligned}
$$

By Lemma 1, we have that the term $\|(\hat{\Sigma}_{JJ} - \Sigma_{JJ})U_J\|$ can be bounded via

$$
\lambda_1 \sqrt{\frac{s\log(p)}{n}}
$$

with probability at least $1 - O(p^{-4})$. Therefore, the bound becomes

$$
\sqrt{\frac{s\log(p)}{n}} \frac{\lambda_1^{3/2}}{\lambda_k(\lambda_k - \lambda_{k+1})} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty} \right),
$$

which is the desired bound.

It remains to show that the integrals tend to zero along the curves $\mathcal{C}_2$ and $\mathcal{C}_4$. Let $\mathcal{I}(z)$ denote the integrand. Then for sufficiently large γ,

$$\begin{aligned}\oint_{\mathcal{C}_2} \|\mathcal{I}(z)\| dz &= \int_a^b \left\| \Lambda^{1/2}(\Lambda - (x+\gamma i)I)^{-1} U_J^\top (\hat{\Sigma}_{JJ} - \Sigma_{JJ})[\tilde{U}_J \tilde{U}_\perp](\hat{\Lambda}_{all} - (x+\gamma i)I)^{-1} \right\| dx \\ &\leq (b-a) \sup_{x \in [a,b]} \left\| \Lambda^{1/2}(\Lambda - (x+\gamma i)I)^{-1} U_J^\top (\hat{\Sigma}_{JJ} - \Sigma_{JJ})[\tilde{U}_J \tilde{U}_\perp](\hat{\Lambda}_{all} - (x+\gamma i)I)^{-1} \right\| \\ &= O(\gamma^{-2}),\end{aligned}$$

which tends to zero as $\gamma \to \infty$. □

B.4 Proof of Lemma 6

First, recall the statement of Lemma 6.

Lemma 6 (Bound on J_2)**.** *The term J_2 satisfies*

$$\begin{aligned}\|U_\perp \Lambda_\perp U_\perp^\top \tilde{U}_J\|_{2\to\infty} &\lesssim \kappa^2 \lambda_{k+1} \sqrt{\frac{k \log(p)}{n}} + \lambda_{k+1}\kappa^3 \frac{s \log(p)}{n} \\ &\lesssim \mathcal{E}_4 \lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Recall the definition of J_2 via

$$J_2 := U_\perp \Lambda_\perp U_\perp^\top \tilde{U}_J.$$

Again $U_\perp$ is the matrix such that the $s \times s$ matrix $[U_J, U_\perp]$ is orthogonal.

Proof of Lemma 6. Define the matrix $E := \hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top$. Note that

$$\tilde{U}_J \Lambda - E\tilde{U}_J = U_J U_J^\top \Sigma_{JJ} U_J U_J^\top \tilde{U}_J.$$

Following Cape et al. (2019a) (see also Xie et al. (2019); Tang et al. (2017); Tang (2018)), by Assumption 4, the spectra of E and Λ are disjoint almost surely, so the matrix $\tilde{U}$ can be expanded as a matrix series (Theorem VII.2.2 in Bhatia (1997)) via

$$\tilde{U}_J = \sum_{m=0}^{\infty} E^m (U_J \Lambda U_J^\top) \tilde{U}_J \Lambda^{-(m+1)}.$$

Therefore,

$$J_2 = U_\perp \Lambda_\perp U_\perp^\top \tilde{U}_J = U_\perp \Lambda_\perp U_\perp^\top E U \Lambda U^\top \tilde{U} \Lambda^{-2} + \sum_{m=2}^{\infty} U_\perp \Lambda_\perp U_\perp^\top E^m U \Lambda U_J^\top \tilde{U}_J \Lambda^{-(m+1)}$$

since the 0-th term cancels off because $U_\perp^\top U_J = 0$. Taking the first term and setting R to be the rest of the series, we have that,

$$\|U_\perp \Lambda_\perp U_\perp^\top \tilde{U}_J \tilde{U}_J^\top\|_{2\to\infty} = \|U_\perp \Lambda_\perp U_\perp^\top E U_J \Lambda U_J^\top \tilde{U} \Lambda^{-2}\|_{2\to\infty} + \|R\|_{2\to\infty}, \tag{18}$$

where R is the residual to be bounded. We first bound the leading term. We have that

$$\|U_\perp \Lambda_\perp U_\perp^\top E U_J \Lambda U_J^\top \tilde{U}_J \Lambda^{-2}\|_{2\to\infty} \leq \|U_\perp \Lambda_\perp U_\perp^\top E U_J\|_{2\to\infty} \lambda_k^{-1} \kappa. \tag{19}$$

We note that since $U_\perp^\top U_J = 0$, then

$$E U_J = (\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top) U_J = (\hat{\Sigma}_{JJ} - \Sigma_{JJ}) U_J.$$

Define $\Sigma_{JJ}^{\perp} := U_{\perp}\Lambda_{\perp}U_{\perp}^{\top}$. In light of the block structure in (7), we see that we can write $\Sigma_{JJ}^{\perp} E U_J$ via the sum of the terms

$$\frac{1}{n}(\Sigma_{JJ}^{\perp})\bigg((\Sigma^{1/2})_{JJ}(Y_J^{\top}Y_J - nI)(\Sigma^{1/2})_{JJ} + \Sigma_{JJ^c}^{1/2} Y_{J^c}^{\top} Y_J (\Sigma^{1/2})_{JJ} \\ + (\Sigma^{1/2})_{JJ} Y_J^{\top} Y_{J^c} (\Sigma_{JJ^c}^{1/2})^{\top} + \Sigma_{JJ^c}^{1/2}(Y_{J^c}^{\top}Y_{J^c} - nI)(\Sigma_{JJ^c}^{1/2})^{\top}\bigg)U_J.$$

Recalling that $(\Sigma_{JJ^c}^{1/2})^{\top} U_J = 0$, this yields the only the terms

$$\frac{1}{n}(\Sigma_{JJ}^{\perp})\bigg((\Sigma^{1/2})_{JJ}(Y_J^{\top}Y_J - nI)(\Sigma^{1/2})_{JJ} + \Sigma_{JJ^c}^{1/2} Y_{J^c}^{\top} Y_J (\Sigma^{1/2})_{JJ}\bigg)U_J = \Sigma_{JJ}^{\perp}(\Sigma^{1/2})_{JJ}\bigg(\frac{Y_J^{\top}Y_J}{n} - I\bigg)U_J\Lambda^{1/2} \\ + \Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2}\frac{Y_{J^c}^{\top}Y_J}{n}U_J\Lambda^{1/2}.$$

Define $A_{3/2} := \Sigma_{JJ}^{\perp}(\Sigma^{1/2})_{JJ}$, which satisfies $\|A_{3/2}\| \leq \sqrt{\lambda_1}\lambda_{k+1}$. In $2 \to \infty$ norm, we have that

$$\|A_{3/2}\bigg(\frac{Y_J^{\top}Y_J}{n} - I\bigg)U_J\Lambda^{1/2}\|_{2\to\infty} \leq \sqrt{k\lambda_1}\max_{i,j}\bigg|\bigg(A_{3/2}\bigg(\frac{Y_J^{\top}Y_J}{n} - I\bigg)U_J\bigg)_{ij}\bigg|.$$

Define the matrix M via $M_{kl} := (A_{3/2})_{ik}U_{lj}$. Fixing i and j, note that we can write the i, j entry above as

$$\begin{aligned}\bigg|\sum_{k,l} M_{kl}\bigg(\frac{1}{n}(\sum_{q=1}^{n}(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk})\bigg)\bigg| &= \frac{1}{n}\bigg|\sum_{q}\sum_{k,l} M_{kl}\bigg(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}\bigg)\bigg| \\ &\leq \frac{1}{n}\sum_{q}\bigg|\sum_{k,l} M_{kl}\bigg(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}\bigg)\bigg| \\ &\leq \max_{q}\bigg|\sum_{k,l} M_{kl}\bigg(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}\bigg)\bigg|,\end{aligned}$$

which is a quadratic form in the random variables Y_{ql} (for fixed q). To bound this, we will apply the Hanson-Wright inequality (Theorem 4 in Appendix C), which requires bounding the Frobenius norm of M. Note that we can bound the Frobenius norm of M via

$$\begin{aligned}\|M\|_F^2 &= \sum_{k,l} M_{kl}^2 \\ &= \sum_{k,l}(A_{3/2})_{ik}^2 U_{lj}^2 \\ &= \|A_{3/2}\|_{2\to\infty}^2 \\ &\leq \bigg(\sqrt{\lambda_1}\lambda_{k+1}\bigg)^2.\end{aligned}$$

Therefore, applying the Hanson-Wright inequality shows that

$$\mathbb{P}\bigg(\bigg|\sum_{k,l} M_{kl}\bigg(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}\bigg)\bigg| > t\bigg) \leq 2\exp\bigg(-c\min\bigg\{\frac{t^2}{\|M\|_F^2}, \frac{t}{\|M\|}\bigg\}\bigg).$$

Set $t := C\sqrt{\frac{\log(s)+\log(k)+5\log(p)}{n}}\sqrt{\lambda_1}\lambda_{k+1}$. Then since $\frac{\log(p)}{n} = o(1)$, we see that with probability at least $1 - O(s^{-1}k^{-1}p^{-5})$ that

$$\bigg|\sum_{k,l} M_{kl}\bigg(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}\bigg)\bigg| \lesssim \sqrt{\lambda_1}\lambda_{k+1}\sqrt{\frac{\log(p)}{n}}.$$

Taking a union bound over all n random variables shows that with probability at least $1 - O(s^{-1}k^{-1}p^{-4})$,

$$\sqrt{k\lambda_1}\left|\left(A_{3/2}\left(\frac{Y_J^\top Y_J}{n} - I\right)U_J\right)_{ij}\right| \lesssim \lambda_1\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}}.$$

Taking a union bound over all s rows and k columns yields that with probability at least $1 - O(p^{-4})$,

$$\|A_{3/2}\left(\frac{Y_J^\top Y_J}{n} - I\right)U_J\Lambda^{1/2}\|_{2\to\infty} \lesssim \lambda_{k+1}\lambda_1\sqrt{\frac{k\log(p)}{n}}. \tag{20}$$

For the other term, proceeding similarly,

$$\begin{aligned}\|\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c}\frac{Y_{J^c}^\top Y_J}{n}U_J\Lambda^{1/2}\|_{2\to\infty} &\leq \sqrt{\lambda_1 k}\max_{i,j}\left|\left((\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c})\frac{Y_{J^c}^\top Y_J}{n}U_J\right)_{ij}\right| \\ &\leq \sqrt{\lambda_1 k}\max_{i,j}\max_q\left|\sum_{k=s+1}^{p}\sum_{l=1}^{s}(\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c})_{ik}Y_{qk}Y_{ql}(U_J)_{lj}\right|.\end{aligned}$$

For fixed q, i, and j, note that k ranges from $s+1$ to p and l ranges from 1 to s, so this is a sum of independent exponential random variables. We will bound these using Bernstein's inequality (Theorem 3 in Appendix C). Note that the ℓ_2 norm of the coefficients is bounded by

$$\sum_{k=s+1}^{p}\sum_{l=1}^{s}(\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c})^2_{ik}(U_J)^2_{lj} \leq \|\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c}\|^2_{2\to\infty} \leq (2\sqrt{\lambda_1}\lambda_{k+1})^2.$$

Similarly,

$$\begin{aligned}\max_{k,l}|(\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c})_{ik}(U_J)_{lj}| &\leq \|U_J\|_{2\to\infty}\max_{i,k}|e_i^\top(\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c})e_k| \\ &\leq 2\|U_J\|_{2\to\infty}\sqrt{\lambda_1}\lambda_{k+1}.\end{aligned}$$

By the generalized Bernstein Inequality (Theorem 3), we have for any fixed i,j, and q that

$$\mathbb{P}\left(\left|\sum_{k=s+1}^{p}\sum_{l=1}^{s}(\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c})_{ik}Y_{qk}Y_{ql}(U_J)_{lj}\right| > t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{(\sqrt{\lambda_1}\lambda_{k+1})^2}, \frac{t}{\|U\|_{2\to\infty}\sqrt{\lambda_1}\lambda_{k+1}}\right)\right].$$

Taking $t = \sqrt{\lambda_1}\lambda_{k+1}\sqrt{\frac{\log(s)+\log(k)5\log(p)}{n}}$ shows that this holds with probability at least $1 - O(s^{-1}k^{-1}p^{-5})$. Taking a union bound over s rows, k columns, and n different random variables shows that with probability at least $1 - O(p^{-4})$ that

$$\begin{aligned}\|\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c}\frac{Y_{J^c}^\top Y_J}{n}U_J\Lambda^{1/2}\|_{2\to\infty} &\leq \sqrt{\lambda_1 k}\max_{i,j}\left|\left((\Sigma^\perp_{JJ}\Sigma^{1/2}_{JJ^c})\frac{Y_{J^c}^\top Y_J}{n}U_J\right)_{ij}\right| \\ &\lesssim \lambda_{k+1}\lambda_1\sqrt{\frac{k\log(p)}{n}}\end{aligned} \tag{21}$$

Combining (21) and (20) with (19) yields that

$$\begin{aligned}\|U_\perp\Lambda_\perp U_\perp^\top EU_J\Lambda U_J^\top\tilde{U}_J\Lambda^{-2}U_J^\top\|_{2\to\infty} &\lesssim \frac{\kappa}{\lambda_k}\|U_\perp\Lambda_\perp U_\perp^\top EU_J\|_{2\to\infty} \\ &\lesssim \frac{\kappa}{\lambda_k}\left(\lambda_1\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}}\right) \\ &\lesssim \kappa^2\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}}.\end{aligned} \tag{22}$$

So what remains is to bound the residual term R in (18). Recall the definition of R via

$$R := \sum_{m=2}^{\infty}U_\perp\Lambda_\perp U_\perp^\top E^m U_J\Lambda U_J^\top\tilde{U}_J\Lambda^{-(m+1)}.$$

We will bound for each m, but for clarity, we will first study the case $m = 2$. We have that

$$\begin{aligned} U_\perp \Lambda_\perp U_\perp^\top E^2 U_J &= U_\perp \Lambda_\perp U_\perp^\top (\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top)(\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top) U_J \\ &= U_\perp \Lambda_\perp U_\perp^\top (\hat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top)(\hat{\Sigma}_{JJ} - \Sigma_{JJ}) U_J \\ &= U_\perp \Lambda_\perp U_\perp (\hat{\Sigma}_{JJ} - \Sigma_{JJ})^2 U_J + (U_\perp \Lambda_\perp U_\perp)^2 (\hat{\Sigma}_{JJ} - \Sigma_{JJ}) U_J. \end{aligned}$$

The first term is readily bounded by Lemma 1, and the second term can be bounded using the techniques in the previous part of the proof of this Lemma.

We now generalize this strategy for each m, by first providing a similar identity to the one above. Define $\Delta := \hat{\Sigma}_{JJ} - \Sigma_{JJ}$. Note that by definition $E = \Delta + U_\perp \Lambda_\perp U_\perp^\top$ and that $EU_J = \Delta U_J$. Then we have that

$$\begin{aligned} U_\perp \Lambda_\perp U_\perp^\top E^m U_J &= U_\perp \Lambda_\perp^\top U_\perp E^{m-1} \Delta U_J \\ &= U_\perp \Lambda_\perp U_\perp^\top E^{m-2} (\Delta + U_\perp \Lambda_\perp U_\perp) \Delta U_J \\ &= U_\perp \Lambda_\perp U_\perp^\top E^{m-2} \Delta U_J + U_\perp \Lambda_\perp U_\perp^\top E^{m-2} U_\perp \Lambda_\perp U_\perp^\top \Delta U_J. \end{aligned} \tag{23}$$

Let $\mathfrak{s}(m)$ be the set of indices such that $s_1 + ... + s_m = m$. Then for all m we have that

$$U_\perp \Lambda_\perp U_\perp^\top E^m U_J = U_\perp \Lambda_\perp U_\perp^\top \left[\sum_{\mathfrak{s}(m)} \Delta^{s_1} (U_\perp \Lambda_\perp U_\perp^\top)^{s_2} \Delta^{s_3} (U_\perp \Lambda_\perp U_\perp^\top)^{s_4} \cdots (U_\perp \Lambda_\perp U_\perp^\top)^{s_{m-1}} \Delta^{s_m} \right] U_J,$$

which is essentially a noncommutative Binomial Theorem.

First, consider the case that $s_1, ..., s_m$ has only single powers of Δ appearing. If Δ appears all the way on the right hand side; that is, $s_m = 1$, then for any integer m_0, we have that

$$\| U_\perp \Lambda_\perp^{m_0} U_\perp^\top \Delta U_J \|_{2\to\infty} \leq C \lambda_{k+1}^{m_0} \left(\lambda_1 \sqrt{\frac{k \log(p))}{n}} \right),$$

with probability at least $1 - O(p^{-4})$ using analogous techniques to the steps leading up to Equation (22) (i.e. the case $m_0 = 1$). If Δ is not on the right hand side, suppose that its index is $s_g = 1$. Then this term is of the form

$$(U_\perp \Lambda_\perp U_\perp^\top)^{1 + s_1 + s_2 + ... + s_{g-1}} \Delta (U_\perp \Lambda_\perp U_\perp^\top)^{s_{g+1} + ... + s_{m_0}} U_J \equiv 0$$

since $U_\perp^\top U_J = 0$. So the only terms that have at most one factor of Δ appearing are those that show up as ΔU_J.

Next, if $s_1, ..., s_m$ is a set of integers and at least two of the terms s_i that appear on the Δ factor are greater than 1, then

$$\| U_\perp \Lambda_\perp U_\perp^\top \Delta^{s_1} (U_\perp \Lambda_\perp U_\perp^\top)^{s_2} \Delta^{s_3} (U_\perp \Lambda_\perp U_\perp^\top)^{s_4} \cdots (U_\perp \Lambda_\perp U_\perp^\top)^{s_{m-1}} \Delta^{s_m} U_J \|_{2\to\infty} \leq \|\Delta\|^2 \lambda_{k+1}^{m-1},$$

provided that $\|\Delta\| < \lambda_{k+1}$, which happens by Assumption 1 and the spectral norm concentration in Lemma 1 with probability at least $1 - O(p^{-4})$. Fix this event. Then for any m, there are at most 2^m ways to select exponents with a power of at least two on the term $\|\Delta\|$. Therefore, this implies that for fixed m

$$\begin{aligned} \| U_\perp \Lambda_\perp U_\perp^\top E^m U_J \|_{2\to\infty} &\leq \| U_\perp \Lambda_\perp^m U_\perp^\top \Delta U_J \| \\ &\quad + \sum_{\{m: \text{exponent on } \|\Delta\| \text{ is at least } 2\}} \| U_\perp \Lambda_\perp U_\perp^\top \Delta^{s_1} \cdots (U_\perp \Lambda_\perp U_\perp^\top)^{s_{m-1}} \Delta^{s_m} U_J \|_{2\to\infty} \\ &\leq C \lambda_{k+1}^m \left(\lambda_1 \sqrt{\frac{k \log(p))}{n}} \right) + 2^m \lambda_{k+1}^{m-1} \|\Delta\|^2. \end{aligned}$$

This bound corresponds to one such m, and hence is its own event. In order to bound for all m, we follow a strategy in Xia and Yuan (2020). Let $\tilde{m} := \lceil \log(p) \rceil$. Then

$$
\begin{aligned}
\Big\| \sum_{m=2}^{\infty} & U_\perp \Lambda_\perp U_\perp^\top E^m U_J \Lambda U_J^\top \tilde{U}_J \Lambda^{-(m+1)} \Big\|_{2\to\infty} \\
&\le \sum_{m=2}^{\infty} \|U_\perp \Lambda_\perp U_\perp^\top E^m U_J\|_{2\to\infty} \frac{\lambda_1}{\lambda_k^{m+1}} \\
&\le \sum_{m=2}^{\tilde{m}} \|U_\perp \Lambda_\perp U_\perp^\top E^m U_J\|_{2\to\infty} \frac{\lambda_1}{\lambda_k^{m+1}} + \sum_{m=\tilde{m}}^{\infty} \|U_\perp \Lambda_\perp U_\perp^\top E^m U_J\|_{2\to\infty} \frac{\lambda_1}{\lambda_k^{m+1}} \\
&\le \sum_{m=2}^{\tilde{m}} \Bigg(C\lambda_{k+1}^m \Bigg(\lambda_1 \sqrt{\frac{k\log(p))}{n}} \Bigg) \frac{\lambda_1}{\lambda_k^{m+1}} \\
&\qquad + \sum_{m=2}^{\tilde{m}} \Big(2^m \lambda_{k+1}^{m-1} \|\Delta\|^2 \Big) \frac{\lambda_1}{\lambda_k^{m+1}} \\
&\qquad + \sum_{m=\tilde{m}}^{\infty} \frac{\lambda_1}{\lambda_k^{m+1}} \|\Delta\| \lambda_{k+1}^{m+1}.
\end{aligned}
$$

Define

$$
\begin{aligned}
\delta_1 &: = C\kappa \Bigg(\lambda_1 \sqrt{\frac{k\log(p))}{n}} \Bigg) \\
\delta_2 &: = \kappa \lambda_k^{-1} \|\Delta\|^2
\end{aligned}
$$

Then the three sums above can be written as

$$
\begin{aligned}
\delta_1 \sum_{m=2}^{\tilde{m}} \frac{\lambda_{k+1}^m}{\lambda_k^m} + \delta_2 \sum_{m=2}^{\tilde{m}} \frac{2^m \lambda_{k+1}^{m-1}}{\lambda_k^{m-1}} + \lambda_1 \|\Delta\| \sum_{m=\tilde{m}}^{\infty} \frac{\lambda_{k+1}^{m+1}}{\lambda_k^{m+1}} &\lesssim \delta_1 \frac{\lambda_{k+1}^2}{\lambda_k^2} + \delta_2 (1+\varepsilon) \frac{\lambda_{k+1}}{\lambda_k} + \lambda_1 \|\Delta\| \Bigg(\frac{\lambda_{k+1}}{\lambda_k} \Bigg)^{\log(p)} \\
&\lesssim \delta_1 \frac{\lambda_{k+1}^2}{\lambda_k^2} + \delta_2 \frac{\lambda_{k+1}}{\lambda_k} + \lambda_1^2 \sqrt{\frac{s\log(p)}{n}} (1-\varepsilon)^{\log(p)}.
\end{aligned}
$$

Here, the penultimate inequality follows from the fact that by Assumption 4, we have that for some $\varepsilon > 1/64$, $2\lambda_{k+1}/\lambda_k < 1-\varepsilon$, and hence the second term's geometric series converges. The final inequality follows from the assumption $\lambda_{k+1}/\lambda_k < (1-\varepsilon)$. Note that this event holds with probability at least $1 - O(\log(p)p^{-4}) \ge 1 - O(p^{-3})$. Noting that

$$
\|\Delta\| \lesssim \lambda_1 \sqrt{\frac{s\log(p)}{n}}
$$

by Lemma 1, we see that the resulting bound for the residual satisfies

$$
\begin{aligned}
\|R\|_{2\to\infty} &\lesssim \delta_1 \big(\frac{\lambda_{k+1}}{\lambda_k} \big)^2 + \delta_2 \frac{\lambda_{k+1}}{\lambda_k} + \lambda_1^2 \sqrt{\frac{s\log(p)}{n}} (1-\varepsilon)^{\log(p)} \\
&\lesssim \big(\frac{\lambda_{k+1}}{\lambda_k} \big)^2 \kappa \Bigg(\lambda_1 \sqrt{\frac{k\log(p))}{n}} \Bigg) + \frac{\lambda_{k+1}}{\lambda_k} \kappa \lambda_k^{-1} \|\Delta\|^2 + \lambda_1^2 \sqrt{\frac{s\log(p)}{n}} (1-\varepsilon)^{\log(p)} \\
&\lesssim \big(\frac{\lambda_{k+1}}{\lambda_k} \big)^2 \kappa \Bigg(\lambda_1 \sqrt{\frac{k\log(p))}{n}} \Bigg) + \frac{\lambda_{k+1}}{\lambda_k} \kappa \lambda_k^{-1} \|\Delta\|^2 \\
&\lesssim \big(\frac{\lambda_{k+1}}{\lambda_k} \big)^2 \kappa \Bigg(\lambda_1 \sqrt{\frac{k\log(p))}{n}} \Bigg) + \frac{\lambda_{k+1}}{\lambda_k} \kappa \lambda_k^{-1} \lambda_1^2 \frac{s\log(p)}{n} \\
&\lesssim \kappa^2 \lambda_{k+1} \sqrt{\frac{k\log(p)}{n}} + \lambda_{k+1} \kappa^3 \frac{s\log(p)}{n}
\end{aligned}
$$

with probability at least $1 - O(p^{-3})$ by the assumption $\varepsilon > \frac{1}{64}$. Combining with our initial bound in (22), we see that

$$\|J_2\|_{2\to\infty} \lesssim \kappa^2 \lambda_{k+1} \sqrt{\frac{k \log(p)}{n}} + \lambda_{k+1} \kappa^3 \frac{s \log(p)}{n}$$

with probability at least $1 - O(p^{-3})$ as desired. □

B.5 Proof of Lemmas 7 and 8

Recall the statement of Lemma 7.

Lemma 7 (The matrix K_1)**.** *The matrix K_1 satisfies*

$$\begin{aligned} \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\tilde{U}\|_{2\to\infty} &\lesssim \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}} \frac{s \log(p)}{n} + \lambda_1 \sqrt{\frac{k \log(p)}{n}} \\ &\lesssim \mathcal{E}_5 \lambda_k \end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Recall K_1 is given by

$$K_1 := \frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\tilde{U}.$$

Proof of Lemma 7. Note that since $U_J U_J^\top + U_\perp U_\perp^\top = I$, we have that

$$\begin{aligned} \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\tilde{U}_J\|_{2\to\infty} &\leq \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_J U_J^\top \tilde{U}_J\|_{2\to\infty} \\ &\quad + \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_\perp U_\perp^\top \tilde{U}_J\|_{2\to\infty} \\ &\leq \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_J\|_{2\to\infty} \|U_J^\top \tilde{U}_J\| \\ &\quad + \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{2\to\infty} \|U_\perp^\top \tilde{U}_J\| \\ &\leq \sqrt{k}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_J\|_{\max} \\ &\quad + \sqrt{s}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{\max} \|U_\perp^\top \tilde{U}_J\|, \quad (24) \end{aligned}$$

We bound each term inside the max norm, using a strategy similar to the beginning of the proof of Lemma 6. For the first term, note that we can write the absolute value of its i, j entry via

$$\begin{aligned} &\left|\frac{1}{n}\sum_q \sum_{k,l} \bigg((\Sigma^{1/2})_{JJ}\bigg)_{ik} (Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql})) \bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{kj}\right| \\ &\qquad \leq \max_q \left|\sum_{k,l} \bigg((\Sigma^{1/2})_{JJ}\bigg)_{ik} (Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql})) \bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{lj}\right|. \end{aligned}$$

We focus on bounding for fixed q. This is a quadratic form in the random variable $\{Y_{qk}\}_{k=1}^s$. Define the matrix M via

$$M_{kl} := \bigg((\Sigma^{1/2})_{JJ}\bigg)_{ik} \bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{lj}.$$

Note that

$$\begin{aligned} \|M\|_F^2 &= \sum_{k,l} \bigg((\Sigma^{1/2})_{JJ}\bigg)_{ik}^2 \bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{lj}^2 \\ &\leq \lambda_1 \|(\Sigma^{1/2})_{JJ}\|_{2\to\infty}^2 \\ &\leq \lambda_1^2. \end{aligned}$$

Therefore, for any fixed q, i, and j, applying the Hanson-Wright inequality (Theorem 4 in Appendix C),

$$\mathbb{P}\bigg(\bigg|\sum_{k,l}\Big((\Sigma^{1/2})_{JJ}\Big)_{ik}(Y_{qk}Y_{ql}-\mathbb{E}Y_{qk}Y_{ql}))\Big((\Sigma^{1/2})_{JJ}U_J\Big)_{lj}\bigg|>t\bigg)\leq 2\exp\bigg(-c\min\bigg\{\frac{t^2}{\lambda_1^2},\frac{t}{\|M\|}\bigg\}\bigg).$$

Setting $t = C\lambda_1\sqrt{\frac{\log(s)+\log(k)+5\log(p)}{n}}$ and taking a union bound for all n random variables shows that with probability at least $1-O(s^{-1}k^{-1}p^{-4})$ that

$$\max_q\bigg|\sum_{k,l}\Big((\Sigma^{1/2})_{JJ}\Big)_{ik}(Y_{qk}Y_{ql}-\mathbb{E}Y_{qk}Y_{ql}))\Big((\Sigma^{1/2})_{JJ}U_J\Big)_{lj}\bigg|\lesssim\lambda_1\sqrt{\frac{\log(p)}{n}}.$$

Therefore, taking a union bound over all s rows and k columns shows that with probability at least $1-O(p^{-4})$ that

$$\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_J\|_{\max}\lesssim\lambda_1\sqrt{\frac{\log(p)}{n}}. \tag{25}$$

The exact same argument yields with the same probability that

$$\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{\max}\lesssim\lambda_1\sqrt{\frac{\log(p)}{n}}. \tag{26}$$

Combining (24) with (25) and (26) yields

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}\tilde{U}_J\|_{2\to\infty}&\leq\sqrt{k}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_J\|_{\max}\\&\quad+\sqrt{s}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{\max}\|U_\perp^\top\tilde{U}_J\|\\&\lesssim\lambda_1\sqrt{\frac{k\log(p)}{n}}+\lambda_1\sqrt{\frac{s\log(p)}{n}}\|U_\perp^\top\tilde{U}_J\|.\end{aligned}$$

So what remains is to bound the term $\|U_\perp^\top\tilde{U}_J\|$, However, we note that this is simply (by a factor of $\sqrt{2}$) the $\sin\Theta$ distance between the subspace $U_JU_J^\top$ and $\tilde{U}_J\tilde{U}_J^\top$ (see Lemma 10 in Appendix C). Therefore, by Proposition 1, we have that this can be bounded by

$$\|U_\perp^\top\tilde{U}_J\|\lesssim\frac{\lambda_1}{\lambda_k-\lambda_{k+1}}\sqrt{\frac{s\log(p)}{n}}.$$

Putting it all together, this yields that with probability at least $1-O(p^{-4})$ that

$$\begin{aligned}\|K_1\|_{2\to\infty}&=\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}\tilde{U}_J\|_{2\to\infty}\\&\lesssim\frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\frac{s\log(p)}{n}+\lambda_1\sqrt{\frac{k\log(p)}{n}},\end{aligned}$$

which is the desired bound. □

Again, we repeat the statement of Lemma 8.

Lemma 8 (The matrix K_2)**.** *The matrix K_2 satisfies*

$$\begin{aligned}\|\frac{1}{n}\Sigma_{JJ^c}^{1/2}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}\tilde{U}\|_{2\to\infty}&\lesssim\lambda_1\sqrt{\frac{k\log(p)}{n}}+\frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\frac{s\log(p)}{n}\\&\lesssim\mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1-O(p^{-4})$.

Recall that

$$K_2 := \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} \tilde{U}_J$$

Proof of Lemma 8. We have that

$$\begin{aligned}
\|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ}\tilde{U}_J\|_{2\to\infty} &\leq \|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_J\|_{2\to\infty} \\
&\quad + \|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_\perp\|_{2\to\infty}\|U_\perp^\top \tilde{U}_J\| \\
&\leq \sqrt{k}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_J\|_{\max} \\
&\quad + \sqrt{s}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_\perp\|_{\max}\|U_\perp^\top \tilde{U}_J\|. \qquad (27)
\end{aligned}$$

We bound each norm inside the max separately. Define the random variable η_{ij} as the i,j entry of the matrix $\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_J$. Then

$$\eta_{ij} = \frac{1}{n}\sum_{q=1}^{n}\sum_{k=1}^{s}\sum_{l=1}^{p-s}[\Sigma^{1/2}_{JJ^c}]_{il}\xi^{(q)}_{s+l,k}\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{kj},$$

where $\xi^{(q)}_{s+l,k} := Y_{q,s+l}Y_{qk}$. Following a strategy similar to the proof of Lemma 6, we have to bound both the maximum and sum of squared ψ_1 norms of the random variable

$$\alpha_{qlj} := \frac{1}{n}[\Sigma^{1/2}_{JJ^c}]_{il}\xi^{(q)}_{s+l,k}\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{kj}.$$

The squared entries satisfy

$$\|\frac{1}{n}[\Sigma^{1/2}_{JJ^c}]_{il}\xi^{(q)}_{s+l,k}\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{kj}\|^2_{\psi_1} \leq \frac{1}{n^2}([\Sigma^{1/2}_{JJ^c}]_{il})^2\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)^2_{kj}.$$

Summing up over q, l, j,

$$\begin{aligned}
\sum_{q=1}^{n}\sum_{k=1}^{s}\sum_{l=1}^{p-s}\|\alpha_{qlj}\|^2_{\psi_1} &\leq \frac{1}{n}\sum_{k=1}^{s}\sum_{l=1}^{p-s}([\Sigma^{1/2}_{JJ^c}]_{il})^2\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)^2_{kj} \\
&\leq \frac{1}{n}\sum_{k=1}^{s}\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)^2_{kj}\|\Sigma^{1/2}_{JJ^c}\|^2_{2\to\infty} \\
&\leq \frac{\lambda_1\|\Sigma^{1/2}_{JJ^c}\|^2_{2\to\infty}}{n}.
\end{aligned}$$

Also,

$$\max_{q,l,j}\|\alpha_{qlj}\|_{\psi_1} \leq \frac{1}{n}\sqrt{\lambda_1}\|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}.$$

By the the Generalized Bernstein inequality (Theorem 3 in Appendix C),

$$\mathbb{P}\bigg(|\eta_{ij}| > t\bigg) \leq 2\exp\bigg(-cn\min\bigg[\frac{t^2}{\lambda_1\|\Sigma^{1/2}_{JJ^c}\|^2_{2\to\infty}}, \frac{t}{\sqrt{\lambda_1}\|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}}\bigg]\bigg).$$

Again taking $t = C\|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}\sqrt{\lambda_1}\sqrt{\frac{\log(s)+\log(k)+4\log(p)}{n}}$ shows that this holds with probability $1 - O(s^{-1}k^{-1}p^{-4})$. Taking a union over all s rows and k columns of the matrix yields that

$$\begin{aligned}
\|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_J\|_{\max} &\lesssim \|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}\sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}} \\
&\lesssim \lambda_1\sqrt{\frac{\log(p)}{n}}.
\end{aligned}$$

Applying precisely the same argument to the other term yields with probability $1 - O(p^{-4})$ that

$$\|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_\perp\|_{\max} \lesssim \lambda_1 \sqrt{\frac{\log(p)}{n}}.$$

Therefore, combining these bounds with the initial bound in (27) and Proposition 1 and the equivalent expressions for the $\sin\Theta$ distances (Lemma 10 in Appendix C), we have that with probability at least $1 - O(p^{-4})$,

$$\begin{aligned}\|K_2\|_{2\to\infty} = \|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} \tilde{U}_J\|_{2\to\infty} &\leq \sqrt{k}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_J\|_{\max} \\ &\quad + \sqrt{s}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} U_\perp\|_{\max} \|U_\perp^\top \tilde{U}_J\| \\ &\lesssim \lambda_1 \sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}} \frac{s\log(p)}{n}\end{aligned}$$

as desired. □

B.6 Proof of Lemma 9

It will be useful to collect some properties of the matrix $\Sigma^{1/2}_{JJ^c}$, which we state as a proposition.

Proposition 2 (Properties of the Matrix $\Sigma^{1/2}_{JJ^c}$)**.** *The matrix $\Sigma^{1/2}_{JJ^c}$ satisfies*

$$\|\Sigma^{1/2}_{JJ^c}\| \leq 2\sqrt{\lambda_1}$$

Furthermore, the left singular subspace of $\Sigma^{1/2}_{JJ^c}$ must contain columns of $U_\perp$.

Proof of Proposition 2. First, we note that

$$\begin{aligned}\|\Sigma^{1/2}_{JJ^c}\| &= \left\| \begin{pmatrix} 0 & (\Sigma^{1/2})_{JJ^c} \\ ((\Sigma^{1/2})_{JJ^c})^\top & 0 \end{pmatrix} \right\| \\ &\leq \|\Sigma\|^{1/2} + \left\| \begin{pmatrix} (\Sigma^{1/2})_{JJ} & 0 \\ 0 & \Sigma^{1/2}_{J^cJ^c} \end{pmatrix} \right\| \\ &\leq 2\sqrt{\lambda_1},\end{aligned}$$

since eigenvalues bound eigenvalues of any principal submatrix. For the second claim, note that

$$\begin{aligned}\Sigma^{1/2} \begin{pmatrix} U_J \\ 0 \end{pmatrix} &= \begin{pmatrix} (\Sigma^{1/2})_{JJ} & (\Sigma^{1/2})_{JJ^c} \\ ((\Sigma^{1/2})_{JJ^c})^\top & \Sigma^{1/2}_{J^cJ^c} \end{pmatrix} \begin{pmatrix} U_J \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} U_J \\ 0 \end{pmatrix} \Lambda^{1/2}.\end{aligned}$$

This shows that the matrix $(\Sigma^{1/2}_{JJ^c})^\top$ satisfies $(\Sigma^{1/2}_{JJ^c})^\top U_J = 0$, so that its null space must contain the space spanned by U_J. However, this also shows that since $(\Sigma^{1/2}_{JJ^c})^\top \in \mathbb{R}^{(p-s)\times s}$, then its rank is at most $s - k$. Hence, define $(\Sigma^{1/2}_{JJ^c})^\top = V_1 D V_2^\top$ as the reduced singular value decomposition of $(\Sigma^{1/2}_{JJ^c})^\top$. Since its rank is at most $s - k$, we have that $V_1 \in \mathbb{O}(p - s, s - k)$, $V_2 \in \mathbb{O}(s, s - k)$, and D is an $s - k \times s - k$ diagonal matrix of singular values.

Since $(\Sigma^{1/2}_{JJ^c})^\top U_J = V_1 D V_2^\top U_J = 0$, the term $V_2 \in \mathbb{O}(s, s - k)$ must span a space perpendicular to U_J. The only matrix up to choice of basis in $\mathbb{O}(s, s - k)$ satisfying $V_2^\top U_J = 0$ is the matrix $U_\perp$, which establishes the second claim. □

Therefore, all this shows that

- The left singular subspace of $\Sigma^{1/2}_{JJ^c}$ contains columns of $U_\perp$;

- Its singular values are all uniformly bounded by $2\sqrt{\lambda_1}$.

We are now prepared to prove Lemma 9.

Lemma 9 (The matrices K_3 and K_4)**.** *The matrices K_3 and K_4 satisfy*

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}Y_J^\top Y_{J^c}(\Sigma^{1/2}_{JJ^c})^\top \tilde{U}_J\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\\ &\lesssim \mathcal{E}_5\lambda_k;\\ \|\frac{1}{n}\Sigma^{1/2}_{JJ^c}(Y_{J^c}^\top Y_{J^c}-nI)(\Sigma^{1/2}_{JJ^c})^\top \tilde{U}\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1-O(p^{-3})$.

Proof of Lemma 9. Let $\Sigma^{1/2}_{JJ^c}$ have singular value decomposition $U_\perp DV^\top$, where $U_\perp\in\mathbb{O}(s,s-k)$, $D_{ii}\geq 0$, $1\leq i\leq s-k$, $V\in\mathbb{O}(p-s,s-k)$. We will show the result for $D_{ii}>0$, though the same proof goes through if $D_{ii}=0$ for some i.

Then the term K_3 satisfies

$$\begin{aligned}\|K_3\|_{2\to\infty} &= \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}(\Sigma^{1/2}_{JJ^c})^\top\tilde{U}_J\|_{2\to\infty}\\ &\leq \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}VDU_\perp^\top\tilde{U}_J\|_{2\to\infty}\\ &\leq \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}VDU_\perp^\top U_\perp\|_{2\to\infty}\|U_\perp^\top\tilde{U}_J\|\\ &\leq \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}V\|_{2\to\infty}\sqrt{\lambda_1}\|U_\perp^\top\tilde{U}_J\|. \qquad (28)\end{aligned}$$

The term $\|U_\perp^\top\tilde{U}_J\|$ can be bounded via Proposition 1 and Lemma 10 in Appendix C. So what remains is to bound the 2 → ∞ norm in (28). Note that the matrix V is of column dimension at most $(s-k)$. Hence, each of the s rows of the matrix $(\Sigma^{1/2})_{JJ}Y_JY_{J^c}V$ is of dimension at most $s-k$.

Following a strategy similar to that in Lemmas 7 and 8, we have that

$$\begin{aligned}\|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}V\|_{2\to\infty} &\leq \sqrt{s-k}\max_{i,j}\left|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}V\right|_{i,j}\\ &\leq \sqrt{s}\max_{i,j}\left|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}V\right|_{i,j}.\end{aligned}$$

By analogous arguments as in Lemma 8, the i,j entry is a sum of independent mean-zero subexponential random variables, each with ψ_1 norm bounded $\frac{1}{n}\sqrt{\lambda_1}$. Therefore, by Bernstein's inequality, any i,j entry is bounded by

$$C\sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}}$$

with probability at most $1-O(p^{-3})$. Combining with Proposition 1, we have the bound

$$\begin{aligned}\|K_3\|_{2\to\infty} &\lesssim \lambda_1\sqrt{\frac{s\log(p)}{n}}\|U_\perp^\top\tilde{U}_J\|\\ &\lesssim \lambda_1\sqrt{\frac{s\log(p)}{n}}\left(\frac{\lambda_1}{\lambda_k-\lambda_{k+1}}\sqrt{\frac{s\log(p)}{n}}\right)\\ &\lesssim \frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\frac{s\log(p)}{n}\end{aligned}$$

as desired.

For the term K_4, we see that

$$\begin{aligned}\|K_4\|_{2\to\infty} &= \|\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)VDU_\perp^\top \tilde{U}_J\|_{2\to\infty} \\ &\leq \|\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)V\|_{2\to\infty}\sqrt{\lambda_1}\|U_\perp^\top \tilde{U}_J\| \\ &\leq \sqrt{s\lambda_1}\|\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)V\|_{\max}\|U_\perp^\top \tilde{U}_J\|. \end{aligned} \tag{29}$$

We will bound the term inside the max norm for fixed i and j. Observe that

$$\begin{aligned}\bigg|\bigg(\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)V\bigg)_{ij}\bigg| &= \max_{i,j}\bigg|\frac{1}{n}\sum_q\sum_{k,l}\bigg(\Sigma_{JJ^c}^{1/2}\bigg)_{ik}(Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql})V_{lj}\bigg| \\ &\leq \max_q\bigg|\sum_{k,l}\bigg(\Sigma_{JJ^c}^{1/2}\bigg)_{ik}(Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql})V_{lj}\bigg|.\end{aligned}$$

We will first bound the term inside the absolute value for fixed q by Hanson-Wright (Theorem 4 in Appendix C). Let M be the matrix defined via

$$M_{kl} := \bigg(\Sigma_{JJ^c}^{1/2}\bigg)_{ik}V_{lj}.$$

Then

$$\|M\|_F^2 = \sum_{k,l}\bigg(\Sigma_{JJ^c}^{1/2}\bigg)_{ik}^2 V_{lj}^2 = \sum_k\bigg(\Sigma_{JJ^c}^{1/2}\bigg)_{ik}^2 \leq \|\Sigma_{JJ^c}^{1/2}\|_{2\to\infty}^2 \leq 4\lambda_1.$$

Therefore, by applying the Hanson-Wright inequality, for any fixed q it holds that

$$\mathbb{P}\bigg(\bigg|\sum_{k,l}\bigg(\Sigma_{JJ^c}^{1/2}\bigg)_{ik}(Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql})V_{lj}\bigg| \geq t\bigg) \leq 2\exp\bigg(-c\min\bigg\{\frac{t^2}{4\lambda_1}, \frac{t}{\|M\|}\bigg\}\bigg).$$

Setting $t = C\sqrt{\lambda_1}\sqrt{\frac{\log(s)+\log(k)+5\log(p)}{n}}$ and taking a union bound over all q random variables shows that for fixed i and j, with probability at least $1 - O(s^{-1}k^{-1}p^{-4})$,

$$\bigg|\bigg(\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)V\bigg)_{ij}\bigg| \lesssim \sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}}.$$

Taking a union bound over s rows and k columns shows that with probability at least $1 - O(p^{-4})$,

$$\|\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)V\|_{\max} \lesssim \sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}}.$$

Therefore, from the initial bound in (29) and Proposition 1,

$$\begin{aligned}\|K_4\|_{2\to\infty} &= \|\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)VDU_\perp^\top \tilde{U}_J\|_{2\to\infty} \\ &\leq \sqrt{s\lambda_1}\|\Sigma_{JJ^c}^{1/2}\bigg(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\bigg)V\|_{\max}\|U_\perp^\top \tilde{U}_J\| \\ &\lesssim \lambda_1\sqrt{\frac{s\log(p)}{n}}\|U_\perp^\top \tilde{U}_J\| \\ &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\end{aligned}$$

as desired. □

C Background Material on Orlicz Norms, Concentration, and Subspace Perturbation

Here we briefly discuss Orlicz ψ_α Norms and Bernstein's inequality for subexponential random variables.

The *Orlicz Norm* of order α for a real-valued random variable X is defined via

$$\|X\|_{\psi_\alpha} := \inf\{t > 0 : \mathbb{E}\exp(|X|^\alpha/t) \leq 1\}.$$

Random variables with finite ψ_2 norm are called *subgaussian* and those with a finite ψ_1 norm are called *subexponential*. Generally speaking, if X is subgaussian, then X^2 is subexponential and $\|X^2\|_{\psi_1} \lesssim \|X\|_{\psi_2}^2$. One also has the "Cauchy-Schwarz" bound $\|XY\|_{\psi_1} \lesssim \|X\|_{\psi_2}\|Y\|_{\psi_2}$ (Vershynin, 2018).

For subexponential random variables, one has the following generalized Bernstein's inequality. See Theorem 2.8.2 in Vershynin (2018) for the proof.

Theorem 3 (Theorem 2.8.2 in Vershynin (2018))**.** *Let $X_1, ..., X_N$ be independent, mean zero subexponential random variables and let $a = (a_i)_{i=1}^N$. Then there exists a universal constant $c > 0$ such that for all $t \geq 0$, we have that*

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_1}$.

We also make use of the Hanson-Wright Inequality. See Theorem 6.2.1 in Vershynin (2018) for the proof.

Theorem 4 (Hanson-Wright Inequality –Theorem 6.2.1 in Vershynin (2018))**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero subgaussian random variables. Let M be some fixed $N \times N$ matrix. Then there exists a universal constant $c > 0$ such that for all $t \geq 0$, we have that*

$$\mathbb{P}\bigg\{\bigg|\sum_{k,l} M_{kl}X_kX_l - \mathbb{E}M_{kl}X_kX_l\bigg| \geq t\bigg\} \leq 2\exp\bigg(-c\min\bigg\{\frac{t^2}{K^4\|M\|_F^2}, \frac{t}{K^2\|M\|}\bigg\}\bigg),$$

where $K = \max_i \|X_i\|_{\psi_2}$.

We also use several notions from subspace perturbation theory. Suppose U and $\hat{U}$ are two $d_1 \times d_2$ matrices with orthonormal columns with $d_2 \leq d_1$. The $\sin\Theta$ distance between the subspaces spanned by U and $\hat{U}$ is defined as follows. Let $I - UU^\top = U_\perp U_\perp^\top$. Then the (spectral) $\sin\Theta$ distance is defined as

$$\|\sin\Theta(U_1, U_2)\| := \|\hat{U}^\top U_\perp\|.$$

Throughout the supplementary material, we use several equivalent terms for the $\sin\Theta$ distance. We present this here as a lemma, the statement of which is slightly modified from Lemma 1 of Cai and Zhang (2018).

Lemma 10 (Modified from Lemma 1 of Cai and Zhang (2018))**.** *The $\sin\Theta$ distance between two matrices satisfies*

$$\|\sin\Theta(\hat{U}, U)\| \leq \inf_{W:WW^\top = I_{d_2}} \|\hat{U} - UW\| \leq \sqrt{2}\|\sin\Theta(\hat{U}, U)\|;$$
$$\|\sin\Theta(\hat{U}, U)\| \leq \|\hat{U}\hat{U}^\top - UU^\top\| \leq 2\|\sin\Theta(\hat{U}, U)\|.$$