How to reduce dimension with PCA and random projections?

Fan Yang, Sifan Liu, Edgar Dobriban, and David P. Woodruff

***Abstract*—In our "big data" age, the size and complexity of data is steadily increasing. Methods for dimension reduction are ever more popular and useful. Two distinct types of dimension reduction are "data-oblivious" methods such as random projections and sketching, and "data-aware" methods such as principal component analysis (PCA). Both have their strengths, such as speed for random projections, and data-adaptivity for PCA. In this work, we study how to combine them to get the best of both. We study "sketch and solve" methods that take a random projection (or sketch) first, and compute PCA after. We compute the performance of several popular sketching methods (random iid projections, random sampling, subsampled Hadamard transform, CountSketch, etc) in a general "signal-plus-noise" (or spiked) data model. Compared to well-known works, our results (1) give asymptotically exact results, and (2) apply when the signal components are only slightly above the noise, but the projection dimension is non-negligible. We also study stronger signals allowing more general covariance structures. We find that (a) signal strength decreases under projection in a delicate way depending on the structure of the data and the sketching method, (b) orthogonal projections are slightly more accurate, (c) randomization does not hurt too much, due to concentration of measure, (d) CountSketch can be somewhat improved by a normalization method. Our results have implications for statistical learning and data analysis. We also illustrate that the results are highly accurate in simulations and in analyzing empirical data.**

***Index Terms*—Dimension reduction, principal component analysis, sketching, random projection, random matrix theory.**

I. Introduction

In our "big data" age, the size and complexity of data is steadily increasing. Methods for data reduction are used ever more commonly. Among these, dimension reduction methods are used to summarize many features into a small set (see see e.g., the reviews [15], [27], [68] and references therein).

Manuscript received April 30, 2020; revised February 7, 2021. The work of Fan Yang was supported in part by the Wharton Dean's Fund for Postdoctoral Research. David Woodruff would like to thank support from National Institute of Health (NIH) grant 5R01 HG 10798-2, Office of Naval Research (ONR) grant N00014-18-1-256, and a Simons Investigator Award. Edgar Dobriban was supported in part by NSF BIGDATA grant IIS 1837992 and NSF TRIPODS award 1934960.

Fan Yang and Edgar Dobriban are with the Wharton Statistics Department, University of Pennsylvania, Philadelphia, PA 19104, USA (e-mail: fyang75@wharton.upenn.edu; dobriban@wharton.upenn.edu).

Sifan Liu is with the Department of Statistics, Stanford University, Stanford, CA 94305, USA (e-mail: sfliu@stanford.edu).

David P. Woodruff is with the Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA (e-mail: dwoodruf@cs.cmu.edu).

Two prominent and very different classes of dimension reduction exist: *data-oblivious* methods such as random projections and sketching (e.g., [50], [78], [100], [104] etc), and *data-aware* methods such as principal component analysis (PCA) (see e.g., the textbooks and reviews [3], [51], [67], [68] and references therein). This is of course just one way to classify the different methods, as there are also linear and nonlinear approaches, etc.

Both data-oblivious and data-aware methods have their strengths. Data-oblivious methods can be very fast and convenient to implement. Data-aware methods on the other hand can better exploit the structure of the data; e.g., PCA can be statistically optimal under certain conditions (see e.g., [3]).

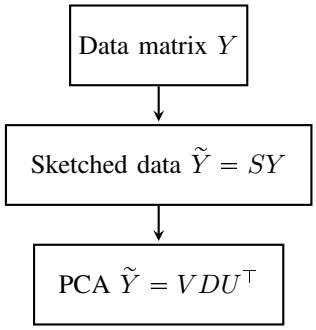

Fig. 1: Flowchart of the algorithm we analyze. We sketch or project the data Y into SY, and then perform PCA.

In this work, we study how to *combine* them to get the best of both. We study "sketch and solve" methods that take a random projection (or sketch) first, and compute PCA after (see Figure 1). Various versions of such algorithms have been proposed [61], [62], [90] (see the related work section). In applied areas, such methods are starting to be used in economics [83], forecasting [92] and genomics [53]. In particular, such algorithms are "state of the art" for dealing with extremely large genomics datasets, where the number of samples (people) is in the thousands to hundreds of thousands, and the number of features (genetic variants, basepairs) is on the order of millions to billions [53].

However, it is not well understood how they perform in all regimes. How can we choose the dimension of sketches? What sketching method—e.g., subsampling or random Gaussian projections—to use? How does their performance depend on the characteristics of the data? Increasing the dimension always increases the accuracy. However, that comes with an increased computational and memory cost. While the existing works do provide theoretical guarantees, they leave some

regimes unstudied [61], [62], [90]. They typically focus on the regime where the signal components are "dominant", and the sketching dimension is very small. As we will see, in a natural model of low-rank data we can get precise results even when the signal components are barely above the noise level, if the sketching dimension is sufficiently large.

In our work we take a systematic approach to this problem. We compute the performance of the most popular sketching methods, such as uniform orthogonal projections, random projections with iid entries, random sampling of the datapoints, subsampled Hadamard transform [1], [91], and count sketch [18], [22]. We work in a general "low-rank signal plus noise" model, sometimes called the "spiked model", which has been widely used to study PCA (e.g., [24], [34], [33], [36], [55], [56], [66], [67], [81], [107] etc). In the spiked model, the entries of the noise matrix are independent random variables, and the signal is an arbitrary low rank matrix that is independent of the noise. This is also a special type of linear factor model.

Compared to well-known classical works [61], [62], [90], our results (1) give asymptotically exact results under more specific assumptions, and (2) apply when the signal components are arbitrarily close to the noise level, provided the sketching dimension is sufficiently large. In particular, a key requirement of our analysis is that the sketching dimension is relatively high, compared to values that practitioners may be interested in for large computational savings. Providing accurate predictions of the behavior of the spikes in the regime where the sketching dimension is much smaller is not covered by our current work and remains important future work. In our analysis, we build on the analysis of recent works in random matrix theory, such as [30], [106]. We also study stronger signals, which allow more general covariance structures. In addition, we illustrate that the results are accurate in simulations and in analyzing empirical data. The computer code reproducing the numerical results in the paper is available from https://github.com/liusf15/sketching-svd.

A. Related work

In this section we review some related work. Due to space limitations, we can only consider the most closely related works. For overviews of sketching and random projection methods, we refer the reader to [40], [62], [78], [100], [104]. A cornerstone result is the Johnson-Lindenstrauss lemma. This states that norms, and thus also relative distances between points, are *approximately* preserved after sketching with an appropriate random matrix S, i.e., $(1-\delta)\|x_i\|^2 \leqslant \|Sx_i\|^2 \leqslant (1+\delta)\|x_i\|^2$ for $x_1, \dots, x_n \in \mathbb{R}^p$. This is further extended to the *subspace embedding property*, that is, for all x in a subspace of relatively small dimension, the norm of x is preserved up to a δ factor. Each projection studied in this paper has the embedding property, and this can be used to derive bounds for the accuracy of PCA. However, our results are much more refined, because they quantify the precise value of the error in an asymptotic setting, while the bounds above are inequalities up to the constant δ.

Compared to well-known classical works [61], [62], [90] on random projection + PCA, our results are in a different data model. Our results give asymptotically exact results when the sample size n and dimension p increase to infinity at the same rate in a spiked model, while the previous results are bounds up to constants. Our results are accurate in simulations and are sharp even when the signal components are only slightly above the noise. In additional related work, Homrighausen and McDonald [63] study the Nyström and column-sampling methods for approximate PCA.

For instance, a typical result, Theorem 1.1 in [62] states that if X is an $n \times p$ data matrix, S is an $r \times n$ sketching matrix with iid standard normal entries, and Q is the $r \times p$ orthogonal matrix projecting into the row space of SX, then

$$\mathbb{E}\|X(I - Q^\top Q)\| \leqslant \left[1 + 4\frac{\sqrt{r}}{r-k-1}\sqrt{\min(n,p)}\right]\sigma_{k+1}.$$

Here σ_{k+1} is the $(k+1)$-st singular value of X. These bounds are sharp if the $(k+1)$-st singular value is small, and r is slightly larger than k. In contrast, our bounds are also applicable to the setting where the $(k+1)$-st singular value of X is only slightly smaller than the k-th one, but we take r to be much larger than k. Thus, our results cover a different regime. In more detail, we consider the regime where n, p, r are all large and proportional to each other, while σ_{k+1} is lower bounded by a constant. In this case, the above bound is on the order of a constant, and hence has limited implications.

Another comparison to prior work is that worst-case bounds for CountSketch are significantly weaker [22], whereas here we find much tighter bounds. For instance, we can effectively show that count-sketch reduces the signal strength by a factor of approximately $\zeta_n(1-\exp(-\zeta_n))$, where $\zeta_n = r/n$ is the ratio of sketched and original sample size. Our bounds more accurately model what is observed in experiments.

However, our results only concern one-step "sketch-and-solve" methods, while there are also other more sophisticated methods. For instance, Frieze et al. [52] proposed randomized SVD using non-uniform row and column sampling. After early works [61], [73], [91] introducing methods based on random projections, Halko et al. [62] developed a unified framework, including iterative algorithms. Woolfe et al. [105] improved the speed via fast matrix multiplications on structured matrices. Musco and Musco [80] proposed a Randomized Block Krylov Iteration methods for fast SVD. Tropp et al. [99] studied the scenario where we can only access the A via a linear map SA. Dasarathy et al. [28] studied how to recover a sparse matrix X from observations AXB. In future work, it will be interesting to extend our approach to those algorithms.

Random projection based approaches have been studied for other problems too, including linear regression [35], [42], [89], [91], ridge regression [20], [74], [76], [102], two sample testing [75], [95], classification [17], convex optimization [86], [87], [88], sparse PCA [54] etc, see [104] for a more comprehensive list of applications.

Compared with prior theoretical work on one-step sketching in linear regression [35], our perspective is similar, in that we want to develop a unified framework to analyze and compare the statistical performance of various sketching methods. In addition, some of the conclusions are also consistent: orthogonal sketches are slightly more accurate, and the subsampled

randomized Hadamard transform is the method of choice overall. However, the similarity stops there. First, this paper is about PCA, a different problem from prior work on linear regression in [35]. Second, this paper considers a theoretical approach leveraging recent results on local laws in random matrix theory [30], [106], while the prior work [35] uses tools such as properties of Wishart matrixes, the generalized Lindeberg principle [19], and liberating sequences from free probability [2]. Thus, the tools are quite different. Finally, an additional difference is that this paper also discusses CountSketch, which has not appeared in [35].

II. Sketching in PCA

In this section we explain our main results. We have an $n \times p$ data matrix Y containing p features of n data points, such as p different measurements on n sensors. We want to perform an approximate PCA of the data. To understand the performance of dimension reduction methods, we assume that the data follows a "low rank signal-plus-noise" or "spiked covariance" matrix model (e.g., [24], [66], [67], [85], [107] etc):

$$Y = WDU^\top + X = \sum_{i=1}^{k} d_i w_i u_i^\top + X.$$

Here $WDU^\top = \sum_{i=1}^{k} d_i w_i u_i^\top$ is the signal component, $\{d_i\}_{1 \leqslant i \leqslant k}$ are the signal strengths (also known as population spikes), and $\{w_i\}_{1 \leqslant i \leqslant k}$ and $\{u_i\}_{1 \leqslant i \leqslant k}$ are the left and right singular vectors of the signals, respectively. They are arranged into the left and right matrices of eigenvectors W and U ($n \times k$ and $p \times k$), and the $k \times k$ diagonal matrix D of population spikes. The matrices U, W are orthogonal: $U^\top U = W^\top W = I_k$. On the other hand, X is the noise component, where the entries x_{ij}, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, are real independent random variables with zero mean and variance $\mathbb{E}|x_{ij}|^2 = n^{-1}$. We assume that any randomness in the signal is independent of the noise matrix X. Other than that, the signal strengths and the singular vectors w_i and u_i can be completely arbitrary. Such signal plus noise or spiked models have been widely studied. When w_i have iid entries, this model can be viewed as a specific *factor model*, and thus has a long history, see e.g., [3], [94], [97].

We consider a setting with large sample size n and dimension p. We place ourselves in a setting where computing a full PCA on Y is too expensive. As an alternative, we are instead interested in PCA on the *sketched* data matrix

$$\widetilde{Y} = SY,$$

where S is an $r \times n$ ($r < n$) random *sketching matrix* that is independent of both the signal and the noise. This can be written as

$$\widetilde{Y} = VDU^\top + SX = \sum_{i=1}^{k} d_i v_i u_i^\top + \widetilde{X},$$

where $V := SW$, $v_i := Sw_i$ and $\widetilde{X} := SX$. A similar spiked separable model has been studied in [30], although the setting there is somewhat different, because the spikes are added to the population covariance matrices instead of to the data. However, we will build on their analysis in our work.

A. Heuristics

Here we explain heuristically what the expected behavior of sketching should be. For simplicity, we consider a one-spiked case, and write the data as $Y = d \cdot wu^\top + X$. Let S be an $r \times n$ partial orthogonal matrix such that $SS^\top = I_r$. Then, we have $SY = d \cdot Swu^\top + SX$. Suppose X has iid Gaussian entries with mean zero and variance n^{-1}. Then $\widetilde{X} = SX$ also has iid Gaussian entries. After projection, the distribution of the noise is unchanged. The low-rank signal changes from $d \cdot wu^\top$ to $\widetilde{d} \cdot vu^\top = d \cdot Swu^\top$, where $\widetilde{d} := d \cdot \|Sw\|$, $v := Sw/\|Sw\|$.

Given the orthogonal invariance of the noise, only the singular values—and not the singular vectors—of the signal govern the behavior of the SVD of the "signal-plus-noise" data. Thus, sketching effectively changes $n \mapsto r$, $d \mapsto d \cdot \|Sw\|$. These fully describe the effect of the projection matrix (which in this case was deterministic). Since $\|Sw\| \leqslant \|S\|\|w\| = 1$, both the sample size and the signal strength get reduced.

However, since we do not know w or $\|Sw\|$, we cannot use the above results to quantify or get insight into the reduction in signal strength. Taking a random S allows us to characterize average behavior of the projections, and thus to get useful predictions. Suppose S is an $r \times n$ random partial orthogonal matrix, i.e., S is uniformly random over the set of matrices such that $SS^\top = I_r$. Then we expect that the norm of $v = Sw$ is $\|v\| \approx (r/n)^{1/2}\|w\|$. This is because we can construct v by randomly rotating w and choosing its first r coordinates. A random rotation makes all coordinates exchangeable, and thus choosing the first r will approximately capture about r/n of the squared norm of w.

Let us write $\xi_n = r/n$ for the reduction in sample size due to sketching. The matrix $\xi_n^{-1/2}\widetilde{X}$ has iid entries of variance r^{-1}. Then the projected matrix SY should be *equivalent to a spiked model with the same spike strength but in a reduced dimension* r:

$$\xi_n^{-1/2}\widetilde{Y} = d \cdot vu^\top + \xi_n^{-1/2}\widetilde{X}. \tag{II.1}$$

Heuristically, after projection into r-dimensional space, both the sample size and the signal strength are reduced by a factor of $\xi_n = r/n$. We will later show rigorously that this is indeed true. For higher dimensional signals, the sketched signal no longer has orthonormal columns, and so the singular values of the signal slightly change. However, since we are dealing with the one-dimensional case in this section, we do not need to worry about this. This shows how taking a random S can simplify our understanding of sketching in PCA. More generally, without Gaussian noise, the randomness in S becomes even more crucial to get interpretable results.

B. Key takeaways

We summarize our key takeaways as follows. Clearly, the signal strength is reduced under projection, and the amount of decrease depends on the type of projection. Moreover:

(i) **Separations between sketching methods**: Our analysis reveals precise separations: the subsampled randomized Hadamard transform (SRHT) and subsampling are—slightly—more accurate than CountSketch and projections with Gaussian or iid entries. For an illustration, we refer the reader

TABLE I: Informal summary of some of our results. For simplicity, suppose we have a single-spike model $Y = d \cdot wu^\top + X$ of size $n \times p$, where $d \cdot wu^\top$ is the signal and X is the noise. We do PCA after sketching on data SY, where S is an $r \times n$ sketching matrix. We show the effective decrease of the signal strength due to sketching. The assumptions needed on X and w depend on the sketching method. The results for iid random S are involved and only presented in the text. Finally, for a more general multi-spike model $Y = \sum_{i=1}^k d_i w_i u_i^\top + X$, we have similar results for the eigenvalues and eigenvectors corresponding to each spike $d_i w_i u_i^\top$.

Assumption on X	Gaussian	independent entries	independent entries	independent entries
Assumption on S	Partial orthonormal	Haar/ Hadamard	Uniform sampling (US)/ CountSketch (CS)	iid random
Assumption on w	Fixed	Fixed	Delocalized	Fixed
Effect on signal	$d \mapsto d \cdot \|Sw\|$	$d \mapsto d \cdot \sqrt{r/n}$	US: $d \mapsto d \cdot \sqrt{r/n}$ CS: $d \mapsto d\sqrt{r/n(1-\exp(-n/r))}$	see Theorem III.2

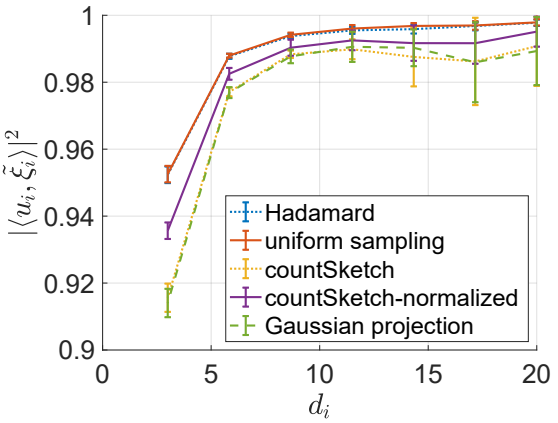

Fig. 2: Comparing SRHT, uniform sampling, CountSketch, and Gaussian projection in a multi-spike model, where $n = 2500$, $p = 800$, $r = 2000$, $k = 7$, and $d_1, \ldots, d_7$ are equally spaced between unity and 20. We show the mean and one standard deviation over 20 Monte Carlo simulations. The normalized CountSketch will be introduced in Section III-E.

to Figure 2, where we plot the overlaps $|\langle u_i, \widetilde{\xi}_i \rangle|^2$ between the population eigenvectors u_i and the sample eigenvectors $\widetilde{\xi}_i$ for a multi-spike model sketched by SRHT, subsampling, CountSketch and projections with iid Gaussian entries. Here a larger overlap indicates that the principal components of the sketched matrix more accurately approximate the principal components of the signal matrix. We remark that the difference between these sketching methods is small in our simulations. However, these minor differences may accumulate in applications, where the sketched matrix can be used repeatedly in iterative algorithms, see e.g., [87]. This may lead to significant differences in downstream applications. The superiority of orthonormal projections is consistent with previous observations in different contexts [35], [71], but our work goes much beyond to include CountSketch and also considers a different problem. However, we point out that we have observed these phenomena in theory and in our numerical simulations only, and not in our experiments with empirical data.

(ii) **Precise quantitative results**: Our results precisely quantify the locations of the sketched spikes. See Table I for an informal summary of some of our results. We show the effective decrease of the signal strengths (i.e., spike strength d) for various sketching methods. However, we will state our formal results in terms of the empirical eigenvalues and eigenvectors, because for some cases (especially for projections S with iid entries), there seems to be no simple way to state them in terms of the decrease in signal strength.

For large signal strengths, we can handle general noise covariance structures, and get simpler results (cf. Section III-F).

(iii) **Additional randomness does not hurt**: A key limitation and drawback of randomized algorithms is that they introduce additional variability in the data. This is an undesireable phenomenon, because the additional variability may lead to vastly different results every time the algorithm is run, and may reduce reproducibility.

In our case, we see that the top eigenvalues and correlations between true and empirical eigenvectors are asymptotically concentrated around definite limits. This means that the additional randomness introduced by the sketching algorithm is relatively limited, for large data sets and for those particular functionals. However, we should still be cautious, in particular about interpreting results obtained from other functionals.

(iv) **Implications for learning**: Our results have implications for statistical learning and signal processing. In particular, by following the methods from [38], [39], they can be used to derive optimal eigenvalue shrinkage estimators for the covariance matrix. Recall that the aforementioned optimal shrinkage operations depend on the overlap between true and empirical eigenvectors. We find those formulas for various sketching methods, and so it becomes possible to use the shrinkage formulas.

Very briefly, Donoho et al. [38] estimate the covariance matrix Σ of the data optimally using eigenvalue shrinkage estimators of the sample covariance matrix. We can replace this with the covariance matrix $\widehat{\Sigma}_s = r^{-1}\widetilde{Y}^\top \widetilde{Y}$ of the sketched data. Let $\widehat{\Sigma}_s = \sum_{i=1}^{\min(r,p)} \widetilde{\lambda}_i \widetilde{\xi}_i \widetilde{\xi}_i^\top$ be the spectral decomposition of $\widehat{\Sigma}_s$, with $\widetilde{\lambda}_i$ sorted in non-increasing order. Then we consider eigenvalue estimators $\widehat{\Sigma}_\eta = \sum_{i=1}^{\min(r,p)} \eta(\widetilde{\lambda}_i) \widetilde{\xi}_i \widetilde{\xi}_i^\top$ for

some fixed shrinker $\eta : \mathbb{R} \to \mathbb{R}$. We evaluate the estimator based on a loss $L(\widehat{\Sigma}_\eta, \Sigma)$, where $\Sigma = I_p + \sum_{i=1}^k d_i^2 u_i u_i^\top$ is the covariance matrix of the original data. For instance, we can have $L(A, B) = \|A - B\|_{\text{op}}$ be the operator norm loss, or $L(A, B) = \|A - B\|_{\text{Fr}}$ be the Frobenius norm loss.

Based on the theory from [38], we can deduce that there is an asymptotically optimal shrinker η for these losses (and a number of others). For instance, for uniform orthogonal random projection, uniform sampling and subsampled randomized Hadamard transform, the optimal shrinkers for operator and Frobenius losses are, respectively,

$$\begin{aligned}\eta_{\text{op}}(x) =& \lambda^{-1}(x^2, r/n),\\ \eta_{\text{Fr}}(x) =& \lambda^{-1}(x^2, r/n) \cdot c^2(\lambda^{-1}(x^2, r/n), r/n)\\ &+ s^2(\lambda^{-1}(x^2, r/n), r/n).\end{aligned}$$

Here λ is the functional inverse of the spike forward map from equation (III.15), and its expression can be found in [38], as well as implemented in software in [32]. Also, c^2 is the cosine forward map from (III.16), and s^2 is the squared sine, defined as $s^2 = 1 - c^2$.

C. How to use our results?

In this section, we give some additional illustration and guidance on how to use our results. Suppose we are interested to compute the SVD or PCA of a massive dataset. Suppose that we are in a setting where we need to use a single machine (possibly after dividing up the data into smaller pieces and distributing them onto different machines—our results can be used at various steps of a broader processing pipeline). Then, a natural approach may be to subsample the $n \times p$ data matrix to $r \leqslant n$ samples. However, this has a chance to miss some datapoints with large entries. Fortunately, there are sketching methods that take linear combinations of each data point, and are thus more likely to pick up these large entries. Which sketching method to use and what projected dimension do we need to get a desired accuracy? What is the appropriate notion of accuracy?

Using our results, we can give some insight into these problems. First, we suggest that we can use a statistical notion of accuracy. Suppose the data is noisy, and we believe that the empirical principal components (PCs) are only estimators of "true" unobserved PCs that one could recover from much more data. Then it makes sense to consider how much sketching reduces the "signal strength" of the PCs in the data at hand. Intuitively, by subsampling r data points out of n, the signal strength should decrease by a factor of r/n. It turns out this intuition is correct, but not at all trivial: it only holds when the "true" principal components are suitably "non-sparse", and requires a somewhat delicate argument. Thus, confirming our intuition, subsampling is only guaranteed to work in a suitable non-sparse setting. However, we show that other orthogonal sketches enjoy the same signal strength reduction, while also working under sparsity. Moreover, some orthogonal sketches, such as randomized Hadamard/Fourier sketches, can be applied in nearly the same time as subsampling. In addition, popular non-orthogonal sketches such as Gaussian projections have strictly worse signal preservation properties than orthogonal ones. Thus, fast orthogonal sketches emerge as the best choice. While this may not be extremely surprising based on prior work, we do believe that it is not commonly discussed in the literature; and in fact we are not aware of a specific work that makes this point for sketched PCA. Furthermore, in Section III-E, we propose a normalized version of CountSketch, which modifies the original CountSketch. Our results suggest that this normalized version has slightly better signal preservation properties than the un-normalized version. Hence our work can be a guide as to when to use the newly proposed normalized CountSketch.

Our results can be also used as a guide to choose the projection dimension. First, suppose we decide that we can tolerate at most a certain factor $f < 1$ (say $f = 1/2$) of decrease in the signal strength. Then, one should use projection dimension r such that $r = fn$ (say $r = n/2$). To compute the estimation error for estimating the true PCs, we can simply use the well-known formulas from spiked covariance models, see Section III-A1. This illustrates how we may use our results. In addition, we believe that we can use our results as a tool to develop and analyze more complicated data analysis methods. However, this is beyond the scope of our current work.

D. Details

Our results require a few more technical assumptions, which are stated in detail in Section A. We use the notion of empirical spectral distribution (ESD) of a matrix M, which is the empirical distribution function of the eigenvalues of M.

In the end, we obtain the following steps for finding the values of the spikes of the sketched matrix $\widetilde{Y} = SY$ (recall S is $r \times n$, with $r \leqslant n$):

(i) For any $x \in [0, \infty)$, we find a fundamental quantity, the pair of weighted *Stieltjes transforms* $(m_{1c}(x), m_{2c}(x))$, as the solution of a certain system of *self-consistent equations* (A.12). Recall that for a distribution F, its Stieltjes transform is defined for any $z \in \mathbb{C}$ away from the support of F as $m_F(z) = \mathbb{E}_{X \sim F}(X - z)^{-1}$ (see e.g., [4], [24], [107]). In our case, (m_{1c}, m_{2c}) are the classical limits of certain weighted Stieltjes transforms (m_1, m_2) of the ESDs of $\widetilde{X}^\top \widetilde{X}$ and $\widetilde{X}\widetilde{X}^\top$ for $\widetilde{X} = SX$ (see Section A-B); and their importance, described below, is in how to use them.

Let us denote by $\gamma_n = p/n$ the aspect ratio, by $\xi_n = r/n \leqslant 1$ the sample size reduction factor, and by $\pi_B := \frac{1}{r}\sum_{i=1}^r \delta_{s_i}$ the empirical spectral distribution of $B = SS^\top$. The self-consistent equation shows that for any $z \in \mathbb{C}_+$ (complex numbers with positive imaginary parts), (m_{1c}, m_{2c}) are determined by the following pair of equations:

$$\begin{aligned} m_{1c}(z) &= \gamma_n \frac{1}{-z\left[1 + m_{2c}(z)\right]},\\ m_{2c}(z) &= \xi_n \int \frac{x}{-z\left[1 + x m_{1c}(z)\right]} \pi_B(\mathrm{d}x). \end{aligned} \tag{II.2}$$

This is a *general Marchenko-Pastur* or *Silverstein* equation; and can also be expressed as a fixed point equation for m_{1c}. It can be solved explicitly in certain special cases. There are also fast numerical solvers available, based on fixed-point methods and ODE solvers, see e.g., [23], [25], [31]. In general, this is

one of the two mathematically challenging parts of the using these steps to characterize the behavior of the spikes after sketching.

(ii) We combine the above quantities into the $2k \times 2k$ *master matrix*

$$M(z) = \begin{pmatrix} M_1(z) & D^{-1} \\ D^{-1} & M_2(z) \end{pmatrix}. \tag{II.3}$$

where M_1 and M_2 are two $k \times k$ matrices defined by

$$M_1(z) := -z^{-1/2}(1 + m_{2c}(z))^{-1} I_k,$$
$$M_2(z) := -z^{-1/2} W^\top S^\top (1 + m_{1c}(z) SS^\top)^{-1} SW.$$

(iii) We solve for the values $x \geqslant 0$ for which the matrix $M(x)$ is singular, i.e.,

$$\det M(x) = 0. \tag{II.4}$$

We call this the *eigenvalue master equation*.

Such determinant equations have appeared in many works in the literature (e.g., [24], [107] and references therein). In general we expect at most k solutions for x. The theory guarantees that these are all possible candidates for the empirical spikes of the sketched data $\widetilde{Y}^\top \widetilde{Y}$. This step turns out to become feasible in several applications due to the randomness in either the sketching matrix S or the signal matrix W. This randomness causes the lower right block to become diagonal, and hence, after rearrangement, the matrix M can be studied as a block matrix with 2×2 blocks.

(iv) To find the angles between the eigenvectors corresponding to an eigenvalue $\widetilde{\lambda}_i$ of $\widetilde{Y}^\top \widetilde{Y}$, again we follow an approach used in many works in the literature (e.g., [24], [107] and references therein). We consider a small contour Γ_i which encloses $\widetilde{\lambda}_i$ (or its classical limit θ_i, as explained below) but no other eigenvalues of $\widetilde{Y}^\top \widetilde{Y}$. The overlap of the corresponding right singular vector $\widetilde{\xi}_i$ with any spike eigenvector u_j of the original data matrix Y is given by the *angle master equation*:

$$|\langle u_j, \widetilde{\xi}_i \rangle|^2 = \frac{1}{2\pi \mathrm{i} (\widetilde{\lambda}_i)^{1/2}} \oint_{\Gamma_i} e_j^\top \mathcal{D}^{-1} M(z)^{-1} \mathcal{D}^{-1} e_j \mathrm{d}z, \tag{II.5}$$

where $\mathcal{D}$ is a $2k \times 2k$ matrix defined by

$$\mathcal{D} := \begin{pmatrix} 0 & D \\ D & 0 \end{pmatrix}.$$

Again, it turns out that in certain cases we can explicitly calculate these integrals.

This finishes the general description of the procedure for finding the sketched spikes. See Section A for details. Next we will go over various popular sketching methods in detail, and show how to use this general procedure.

III. Asymptotics, types of random projections

In this section, we discuss the asymptotic model we consider, we go over the various types of random projections, and explain the behavior of the sketched eigenvalues and eigenvectors. Fix any signal strengths $d_1 > d_2 > \cdots > d_k > 0$. We consider an asymptotic setting where the sample size $n \to \infty$, and at the same time the data dimension $p \to \infty$ and the sketching dimension $r \to \infty$, such that $\gamma_n := p/n \to \gamma > 0$ and $\xi_n := r/n \to \xi > 0$. The setting $p/n \to \gamma > 0$ is common in high-dimensional statistics and random matrix theory [79], [4], [107], while $r/n \to \xi > 0$ has been used to study sketching in [35], [71], etc. In this setting, many of the objects we consider will depend on n,—e.g., $Y = Y_n$ depends on n—but we will drop this dependence for simplicity of notation. Moreover, we will emphasize if certain quantities do not depend on n. For instance, the signal strengths $d_1 > d_2 > \cdots > d_k > 0$ are fixed and will not depend on it.

Without projections, when $S = I_n$, it is well-known that a signal of strength d_i leads to an *outlier* if and only if $d_i^2 > \sqrt{\gamma_n}$ [6], [7]. Here outliers are the eigenvalues of the sample covariance matrix separated and above the "bulk" of the noise eigenvalues, which in the limit of large data is described by a standard Marchenko-Pastur distribution [4], [79]. Moreover, the i-th spiked sample eigenvalue converges to its "classical value"

$$\widetilde{\lambda}_i \to_P \lambda(d_i^2, \gamma) := \left(1 + d_i^2\right)\left(\frac{\gamma}{d_i^2} + 1\right), \tag{III.1}$$

if $\gamma_n \to \gamma$, see e.g., [6], [7]. Here "$\to_P$" denotes convergence in probability. The map $\ell \to \lambda(\ell, \gamma)$ between the population and sample spikes is sometimes referred to as the spiked forward map. The overlaps between population and sample eigenvectors converge to (e.g., [10])

$$|\langle u_j, \widetilde{\xi}_i \rangle|^2 \to_P \delta_{ij} c^2(d_i^2, \gamma) := \delta_{ij} \frac{1 - \frac{\gamma}{d_i^4}}{1 + \frac{\gamma}{d_i^2}}, \tag{III.2}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. The expression $c^2(d^2, \gamma)$ may be referred to as the squared cosine forward map, giving the asymptotic squared cosines between the population and sample eigenvectors. Now we consider several choices of S, and compare the corresponding results to the above formulas. We restrict to a certain high probability event Ω (given formally in (A.25)), where the so-called "local law" holds, and certain empirical quantities are close to their population verions. So "with high probability" means "with high probability on Ω".

A. Uniform orthogonal random projections

We take S to be $r \times n$ partial orthonormal, so that $SS^\top = I_r$.

1) Results known from prior work, Gaussian data: There are a few results that can be readily deduced from known work. They are not our main point (as they are limited to Gaussian data); and our main results can handle much more general data distributions and sketching distributions in a unified framework. However, as they are not explicitly available in prior work, we present them here for the reader's convenience.

When the noise X is has iid Gaussian entries, we have seen in Section II-A that for fixed dimensions n and p, the data $Y = X + WDU^\top$ transforms into $SY = SX + (SW)DU^\top$, with the distribution of SX still being Gaussian. The signals are transformed into $(SW)DU^\top$, and we let its SVD be $\widetilde{W}\widetilde{D}\widetilde{U}^\top$. As is well known from the classical theory of spiked models, [6], the singular values of the signal control the behavior the

data SVD. This shows that we are in a new spiked model with new signal strengths $\widetilde{D}$, which can be checked to have been reduced compared to D.

If in addition we assume that S is distributed uniformly over the Stiefel manifold of partial orthonormal matrices, then it is not hard to check that $\widetilde{D} \approx (r/n)^{1/2} D$, so the signal reduces by a factor of r/n. In addition, the sample size is also reduced. To quantify the change of the outlier eigenvalues and eigenvectors, we recall that after scaling by $\xi_n^{-1/2}$, SY is equivalent to the model in (II.1), which has the same spike strength in a reduced dimension r. In this model, we have that the aspect ratio changes as $\gamma_n \mapsto \frac{p}{r} = \frac{\gamma_n}{\xi_n}$. Thus by (III.1) and (III.2), if $\gamma_n \to \gamma$ and $\xi_n \to \xi$, we have that

$$\begin{aligned} \widetilde{\lambda}_i \to_P \theta_i &:= \xi \left(1 + d_i^2\right) \left(\frac{\gamma/\xi}{d_i^2} + 1\right) \\ &= \left(1 + d_i^2\right) \left(\frac{\gamma}{d_i^2} + \xi\right), \end{aligned} \tag{III.3}$$

and

$$|\langle u_j, \widetilde{\xi}_i \rangle|^2 \to_P \delta_{ij} \frac{1 - \frac{\gamma/\xi}{d_i^4}}{1 + \frac{\gamma/\xi}{d_i^2}} = \delta_{ij} \frac{\xi - \frac{\gamma}{d_i^4}}{\xi + \frac{\gamma}{d_i^2}}, \tag{III.4}$$

for the eigenvalues and eigenvectors of SY. One can compare them to the results in (III.1) and (III.2), and see how the location of the spikes decreases.

The same logic applies to all distributions of partial orthonormal sketching matrices S and all signal matrices W for which $(SW)^\top SW \approx r/n \cdot W^\top W = r/n \cdot I_k$. We will discuss this for each case separately.

2) New results, general data: When the distribution of the data is general, the above direct argument cannot be used. We will instead use our general framework. We assume S is distributed uniformly over the Stiefel manifold of partial orthonormal matrices. Then the self-consistent equation for the Stieltjes transforms (II.2) becomes

$$\begin{aligned} m_{1c}(z) &= \gamma_n \frac{1}{-z\left[1 + m_{2c}(z)\right]}, \\ m_{2c}(z) &= \xi_n \frac{1}{-z\left[1 + m_{1c}(z)\right]}, \end{aligned} \tag{III.5}$$

from which we can obtain the following equation for $m_{2c}(z)$:

$$z m_{2c}^2 + (z - \gamma_n + \xi_n) m_{2c} + \xi_n = 0. \tag{III.6}$$

This equation has a unique solution with non-negative imaginary part for $z \in \mathbb{C}_+$, that is,

$$m_{2c}(z) = \frac{-(z - \gamma_n + \xi_n) + \sqrt{(z - \lambda_+)(z - \lambda_-)}}{2z}, \tag{III.7}$$

where

$$\lambda_\pm = (\sqrt{\gamma_n} \pm \sqrt{\xi_n})^2.$$

Moreover, m_{2c} is injective on the $\{z : \operatorname{Re} z > \lambda_+\}$, and we denote its inverse function as g_{2c}, which takes the form

$$g_{2c}(m) = \frac{\gamma_n}{1 + m} - \frac{\xi_n}{m}. \tag{III.8}$$

Given m_{1c} and m_{2c}, we now study the master matrix $M(z)$ in (II.3). We can write the matrix W of eigenvectors as

$$W = \mathbb{W} \begin{pmatrix} I_k \\ 0 \end{pmatrix},$$

where $\mathbb{W}$ is an $n \times n$ orthogonal matrix. Now recall that $V = SW$. Since the distribution of S is rotationally invariant, we have

$$V^\top V = \begin{pmatrix} I_k, 0 \end{pmatrix} \widehat{S}^\top \widehat{S} \begin{pmatrix} I_k \\ 0 \end{pmatrix},$$

where $\widehat{S} = S\mathbb{W}$ is, like S, also an $r \times n$ uniformly distributed partial orthogonal matrix. We claim that

$$V^\top V = \xi_n I_k + \mathrm{o}(1) \quad \text{in probability.} \tag{III.9}$$

This can be easily verified by a simple variance calculation using exchangeability of the rows or columns of $\widehat{S}$; see Appendix B. With (III.9), the eigenvalue master equation (II.4) becomes

$$\det \begin{pmatrix} -\frac{x^{-1/2}}{1 + m_{2c}(x)} I_k & D^{-1} \\ D^{-1} & x^{1/2} m_{2c}(x) I_k \end{pmatrix} = \mathrm{o}(1) \tag{III.10}$$

in probability. Ignoring the small (random) error on the right-hand side, the above matrix equation holds if and only for some $i = 1, \ldots, k$,

$$\begin{aligned} &\det \begin{pmatrix} -\frac{x^{-1/2}}{1 + m_{2c}(x)} & d_i^{-1} \\ d_i^{-1} & x^{1/2} m_{2c}(x) \end{pmatrix} = 0 \\ &\Leftrightarrow m_{2c}(x) = -\frac{1}{1 + d_i^2} \Leftrightarrow x = g_{2c}\left(-\frac{1}{1 + d_i^2}\right), \end{aligned} \tag{III.11}$$

where in the second step we used that g_{2c} is the inverse function of m_{2c}. This gives an equation for any potential outlier x. However, in order to have an outlier, we need to have that

$$-\frac{1}{1 + d_i^2} > m_{2c}(\lambda_+) = \frac{-1}{1 + \sqrt{\frac{\gamma_n}{\xi_n}}} \Leftrightarrow d_i^2 > \sqrt{\frac{\gamma_n}{\xi_n}}. \tag{III.12}$$

This is because the Stieltjes transform m_{2c} is increasing on $[\lambda_+, \infty)$ outside of the bulk of eigenvalues, and so having $\lambda_+ < x$ is equivalent to $m_{2c}(\lambda_+) < m_{2c}(x) = -1/(1 + d_i^2)$. Then we use the known formula for $m_{2c}(\lambda_+)$ given in (III.7). Using (III.8) and (III.11), we obtain that the classical location for the outlier caused by d_i is

$$\theta_i = \left(1 + d_i^2\right) \left(\frac{\gamma_n}{d_i^2} + \xi_n\right). \tag{III.13}$$

This formula is very similar to the well known one for the location of the empirical spike in standard spiked models, presented above, and described in [6], [7]. However, we have a different setting in this paper, and so the formula does not follow from the classical one.

Next we turn to finding the formula for the angle between projected and true spikes. Using (II.5), we obtain that $|\langle u_j, \widetilde{\xi}_i \rangle|^2 = \mathrm{o}(1)$ if $j \neq i$. If $j = i$, the inverse of $M(z)$ is also a block matrix with 2×2 blocks, and so its (i, i)-th entry

is approximately the appropriate entry of the inverse of the 2×2 block it belongs to. Therefore, we have in probability,

$$
\begin{aligned}
|\langle u_i, \widetilde{\xi}_i \rangle|^2 &= \mathrm{o}(1) + \frac{1}{2\pi \mathrm{i} \sqrt{\theta_i}} \oint_{\Gamma_i} (0, d_i^{-1}) \\
&\times \begin{pmatrix} -z^{-1/2} \left(1 + m_{2c}(z)\right)^{-1} & d_i^{-1} \\ d_i^{-1} & z^{1/2} m_{2c}(z) \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ d_i^{-1} \end{pmatrix} \mathrm{d}z \\
&= \frac{1}{2\pi \mathrm{i} \theta_i \left(1 + d_i^2\right)} \oint_{\Gamma_i} \frac{1}{m_{2c}(z) + (1 + d_i^2)^{-1}} \mathrm{d}z + \mathrm{o}(1) \\
&= \frac{1}{\theta_i \left(1 + d_i^2\right) m_{2c}'(\theta_i)} + \mathrm{o}(1) = \frac{g_{2c}'(-(1 + d_i^2)^{-1})}{\theta_i \left(1 + d_i^2\right)} + \mathrm{o}(1) \\
&= \frac{\xi_n - \frac{\gamma_n}{d_i^4}}{\xi_n + \frac{\gamma_n}{d_i^2}} + \mathrm{o}(1). \qquad \text{(III.14)}
\end{aligned}
$$

See Figure 3 for simulation results illustrating the accuracy of the formulas (III.13) and (III.14). We generate $Y = WDU^\top + X$ with iid entries $x_{ij} \sim \mathrm{Unif}(-\sqrt{3/n}, \sqrt{3/n})$ (so x_{ij} has variance $1/n$), and take the rank $k = 1$ with W, U being independent uniformly distributed partial orthogonal random matrices; or, since $k = 1$, random orthogonal vectors. We vary the value for the spike d_1, and compute a random projection $\widetilde{Y} = SY$ with a uniformly random partial orthogonal matrix S. We then compute its SVD and find its first eigenvalue and eigenvector. We compare them to the theoretical formulas above. See the caption to Figure 3 for more details. In Section IV-C, we will provide simulation results where the rank k is greater than one.

Now we state the above results as the following theorem. We shall prove it rigorously in Appendix B.

Theorem III.1 (Uniform orthogonal random projection)**.** *Consider the $r \times p$ sketched data matrix $\widetilde{Y} = SY$, where S is $r \times n$ partial orthonormal, distributed uniformly over the Stiefel manifold of partial orthonormal matrices. Also, $X = (x_{ij})$ is an $n \times p$ random matrix where the entries x_{ij} are real independent random variables with mean zero, variance n^{-1}, with their higher moments bounded as in* (A.8)*. Let the number of signals k be a finite fixed integer and the strengths $d_1 > d_2 > \cdots > d_k > 0$ be fixed constants; $\{u_i\}_{1 \leqslant i \leqslant k}$ and $\{w_i\}_{1 \leqslant i \leqslant k}$ be deterministic sets of orthonormal unit vectors in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively. Let $\gamma_n := p/n \to \gamma$ and $\xi_n := r/n \to \xi$ as $n \to \infty$. Then for any $1 \leqslant i \leqslant k$, if $d_i > \sqrt{\gamma/\xi}$, we have*

$$
\widetilde{\lambda}_i \to_P \left(1 + d_i^2\right) \left(\frac{\gamma}{d_i^2} + \xi\right), \qquad \text{(III.15)}
$$

and

$$
|\langle u_j, \widetilde{\xi}_i \rangle|^2 \to_P \delta_{ij} \frac{\xi - \frac{\gamma}{d_i^4}}{\xi + \frac{\gamma}{d_i^2}}. \qquad \text{(III.16)}
$$

Otherwise, if $d_i \leqslant \sqrt{\gamma/\xi}$, we have

$$
\widetilde{\lambda}_i \to_P \lambda_+, \qquad \text{(III.17)}
$$

and

$$
|\langle u, \widetilde{\xi}_i \rangle|^2 \to_P 0 \qquad \text{(III.18)}
$$

for any sequence of deterministic unit vectors $u := u_p \in \mathbb{R}^p$.

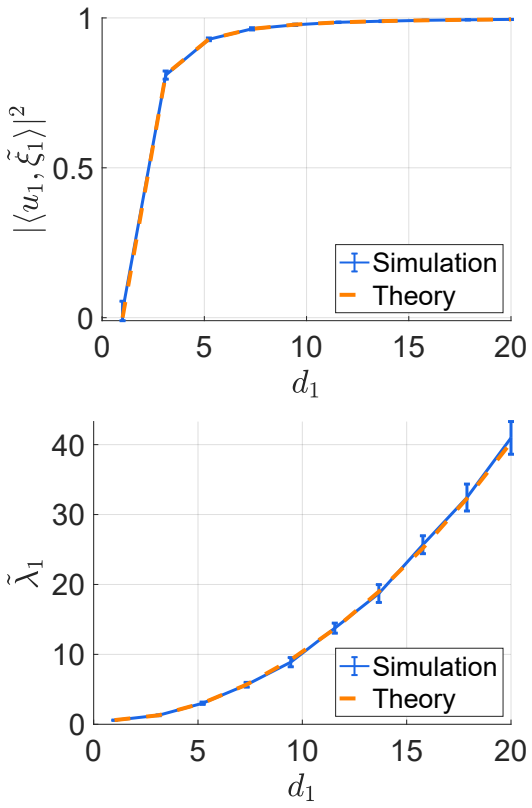

Fig. 3: Checking the accuracy of the spiked eigenvalue and eigenvector formulas for orthogonal projections. We show results with $n = 4000$, $p = 800$, $r = 400$ and $k = 1$, and we vary the signal strength d_1 between unity and 20. The entries x_{ij} are iid random variables sampled from $\mathrm{Unif}(-\sqrt{3/n}, \sqrt{3/n})$. We show the mean and one standard deviation over 20 Monte Carlo simulations (the SD is very small). Top: d_1 against the overlap between the spiked population eigenvector and the sample eigenvector after sketching. Bottom: d_1 against $\widetilde{\lambda}_1$, the first sample eigenvalue after sketching. The theoretical and empirical results agree well.

This result is consistent with our heuristics that the sketched spiked model should be *equivalent to a spiked model* (II.1) *with the same spike strengh but in a reduced dimension r*. Moreover, the theory is also consistent with the results readily deduced from prior work for Gaussian data presented in (III.3) and (III.4).

B. Projections with iid entries

Now we pick S to be an $r \times n$ random matrix with iid entries of zero mean, variance n^{-1}, and with bounded moments, as in (A.8). In particular, S can be a random Gaussian projection if its entries are iid Gaussian. Then $B = SS^\top$ is a sample covariance matrix with identity population covariance. Before giving the main result, Theorem III.2, we first introduce the notations that are used in its statement.

First, we define two functions m_{1c}^S and m_{2c}^S, which are the Stieltjes transforms of the well-known Marchenko-Pastur law

(m_{1c}^S is for $SS^\top$ and m_{2c}^S is for $S^\top S$):

$$m_{1c}^S(z) = \frac{-(z-1+\xi)+\sqrt{(z-\lambda_+^S)(z-\lambda_-^S)}}{2z\xi},$$
$$m_{2c}^S(z) = \frac{-(z+1-\xi)+\sqrt{(z-\lambda_+^S)(z-\lambda_-^S)}}{2z},$$

where $\lambda_\pm^S$ are the edges of the support of the MP law, $\lambda_\pm^S = (1\pm\sqrt{\xi})^2$. Then g_{1c} is defined as

$$g_{1c}(m) = -\frac{\gamma}{m} + \frac{\xi}{m}\left(1 - \frac{1}{m}m_{1c}^S(-m^{-1})\right).$$

In fact, g_{1c} is the inverse function of $m_{1c}(z)$, which is the unique solution to the cubic equation

$$z^2 m_{1c}^3 - z(1+\xi-2\gamma)m_{1c}^2 - \left[z+(1-\gamma)(\gamma-\xi)\right]m_{1c} - \gamma = 0,$$

that satisfies $\operatorname{Im} m_{1c}(z) > 0$ for any z with $\operatorname{Im} z > 0$. It is possible to give an explicit expression of $m_{1c}(z)$ using the formulas for the roots of cubic equations, but we do not state it here. Taking the limit as $\operatorname{Im} z \downarrow 0$, it follows from our results that we obtaina continuous function $\rho_{1c}(x) := \lim_{\eta\downarrow 0}\pi^{-1}\operatorname{Im} m_{1c}(x+\mathrm{i}\eta)$. The function $\xi^{-1}\rho_{1c}$ is a probability density function compactly supported on $\mathbb{R}_+ := [0,\infty)$, and we denote the rightmost edge of its support by λ_+ following the convention in random matrix theory. Then we define $\alpha_i \equiv \alpha(d_i)$ as

$$\alpha_i \equiv \alpha(d_i) := -\frac{\gamma d_i^{-2}}{\left(1+\gamma d_i^{-2}\right)\left(\xi+\gamma d_i^{-2}\right)},$$

and $d_c > 0$ is defined as the unique solution (which we show is well-defined) to the equation

$$\alpha(d_c) = m_{1c}(\lambda_+).$$

We have the following result for sketching with iid projection, which will be proved in Appendix C.

Theorem III.2 (Random projection with iid entries)**.** *Suppose that the assumptions in Theorem III.1 hold except that S is an $r\times n$ random sketching matrix whose entries are independent random variables of zero mean, variance n^{-1}, and with bounded moments, as in* (A.8)*. Then for any $1\leqslant i\leqslant k$, if $d_i > d_c$, we have*

$$\widetilde{\lambda}_i \to_P \theta_i := g_{1c}(\alpha_i), \tag{III.19}$$

and

$$|\langle u_j, \widetilde{\xi}_i\rangle|^2 \to_P \delta_{ij}\frac{\alpha_i^2}{d_i^2}\frac{g_{1c}'(\alpha_i)}{\left[(m_{2c}^S)'(-\alpha_i^{-1})\right]\alpha_i^{-2} - (1+\gamma d_i^{-2})}. \tag{III.20}$$

Otherwise, if $d_i \leqslant d_c$, then (III.17) *and* (III.18) *hold.*

We can get explicit expressions for the right-hand sides of (III.19) and (III.20) using the formulas for m_{1c}^S, m_{2c}^S and g_{1c}. Since they are very complicated, we do not state them here.

Algorithmically, given γ_n, ξ_n, d_i^2, we find the location of the spike after random projection by calculating the value of (III.19) using the above formula for g_{1c} (which involves m_{1c}^S).

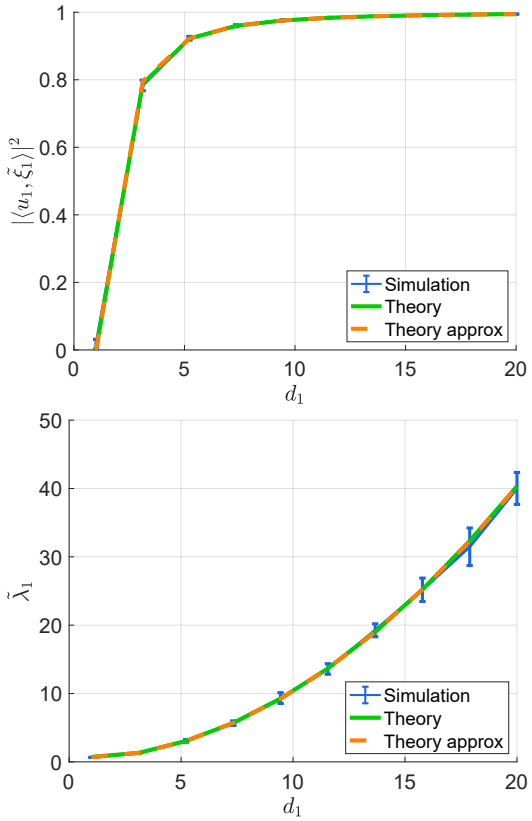

Fig. 4: Checking the accuracy of the spiked eigenvalue and eigenvector formulas for Gaussian projections. We follow the protocol from the experiment in Figure 3.

See Figure 4 for simulations checking the accuracy of these results.

Now we compare sketches with iid random entries with uniform random projections discussed in the previous section. The explicit expressions for θ_i and $|\langle u_i, \widetilde{\xi}_i\rangle|^2$ are pretty cumbersome. To simplify the expressions, we consider the large signal case where d_i is a large constant, and develop asymptotic expressions of θ_i and $|\langle u_i, \widetilde{\xi}_i\rangle|^2$ in terms of d_i^{-2}. Through direct calculation, we find that

$$\begin{aligned}
(m_{1c}^S)'(-\alpha_i^{-1}) &= \alpha_i^2\left(1-2\alpha_i+(3+3\xi)\alpha_i^2+\mathrm{O}(\alpha_i^3)\right),\\
(m_{2c}^S)'(-\alpha_i^{-1}) &= \alpha_i^2\left(1-2\xi\alpha_i+(3+3\xi)\xi_n\alpha_i^2+\mathrm{O}(\alpha_i^3)\right),\\
m_{1c}^S(-\alpha_i^{-1}) &= \alpha_i\left(1-\alpha_i+(\xi+1)\alpha_i^2+\mathrm{O}(\alpha_i^3)\right),\\
g_{1c}'(\alpha_i) &= \frac{\gamma}{\alpha_i^2} - \xi(\xi+1) + \mathrm{O}(\alpha_i).
\end{aligned}$$

Plugging these into (III.19) and (III.20), we obtain that

$$\theta_i = \xi d_i^2 + (\xi\gamma+\gamma+\xi) + (\gamma+\xi+1)\gamma d_i^{-2} + \mathrm{O}(d_i^{-4}), \tag{III.21}$$

and

$$\begin{aligned}
|\langle u_i, \widetilde{\xi}_i\rangle|^2 &\to_P \frac{\gamma-\xi(1+\xi)\alpha_i^2+\mathrm{O}(\alpha_i^3)}{d_i^2}\\
&\quad\times\frac{1}{-2\xi\alpha_i+(3+3\xi)\xi\alpha_i^2+\mathrm{O}(\alpha_i^3)-\gamma d_i^{-2}}\\
&= \frac{\xi-\frac{(1+\xi)\gamma}{d_i^4}+\mathrm{O}(d_i^{-6})}{\xi+\frac{(1+\xi)\gamma}{d_i^2}+\mathrm{O}(d_i^{-4})}.
\end{aligned} \tag{III.22}$$

Compared with (III.16), one can see that, at least in the large signal regime, the correlation (III.22) is smaller, and thus worse than random uniform projection. Moreover, we have the simple relation

$$\frac{(1-|\langle u_i,\widetilde{\xi}_i\rangle|^2)_{\text{iid projection}}}{(1-|\langle u_i,\widetilde{\xi}_i\rangle|^2)_{\text{uniform projection}}}=1+\xi+\mathrm{O}(d_i^{-2})$$

for these two cases, where the notations are self-explanatory.

One can see Figure 4 for simulations checking the accuracy of the results (III.21) and (III.22). Surprisingly, even for small d_i, they are already sufficiently precise.

C. Uniform random sampling

Next, we take S to be an $n\times n$ diagonal sampling matrix, where the entries S_{ii} are iid with

$$S_{ii}=\varepsilon_i, \tag{III.23}$$

where $\varepsilon_i \sim Bernoulli(r/n)$. This is closely related to sampling r out of n datapoints uniformly at random, as for large r and n the number sampled concentrates around $r\pm\mathrm{O}(\sqrt{r(1-r/n)})\approx r$. Then we find the following result.

Theorem III.3 (Uniform random sampling)**.** *Suppose that the assumptions in Theorem III.1 hold, but S is a random sampling matrix as in* (III.23)*. Assume that the vectors w_i are delocalized in the following sense:*

$$\max_{1\leqslant i\leqslant k}\|w_i\|_\infty\to 0 \quad as \quad n\to\infty. \tag{III.24}$$

Then the results (III.15)-(III.18) *hold.*

The proof of Theorem III.3 is a minor modification of the one for Theorem III.1 in Appendix B, and we highlight the differences in Section D. See Figure 5 for experimental results supporting these theoretical results.

We remark that the delocalization condition (III.24) is necessary for uniform random sampling. Indeed, suppose for instance that w_i only contains one non-zero entry. Then uniform random sampling has a positive probability of missing this entry, so that $Sw_i=0$. In this case, the principal components of the sketched matrix SY will deviate greatly from those of Y.

When w_i are not delocalized, it is more natural to use non-uniform sampling methods. We refer the reader to Section V for a discussion of some more advanced sampling methods.

Gaussian data. For the special case of Gaussian data, recall from Section III-A1 that we can readily deduce Theorem III.3 from known spiked model results if we can show $W^\top S^\top SW=\sum_i\varepsilon_i w[i]^\top w[i]\approx r/n\cdot I_k$, where $w[i]$ are the rows of W. This follows from standard matrix concentration results (e.g., [98], [101] etc). Since this is not our main point, we will not elaborate it in more detail.

D. Randomized Hadamard sampling

We consider the subsampled randomized Hadamard transform. Define the $n\times n$ subsampled randomized Hadamard matrix as

$$S=\frac{1}{\sqrt{n}}B_rHD, \tag{III.25}$$

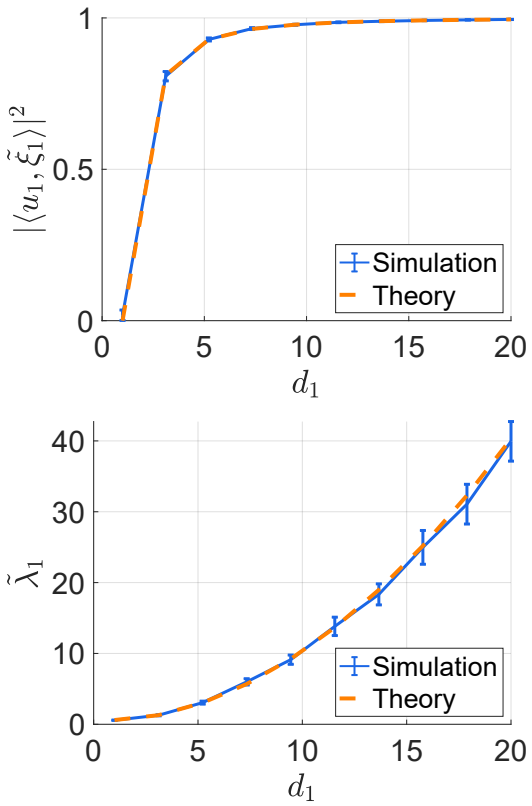

Fig. 5: Checking the accuracy of the spiked eigenvalue and eigenvector formulas for uniform random sampling. We follow the protocol from the experiment in Figure 3.

where B_r is a diagonal sampling matrix with iid $Bernoulli(r/n)$ diagonal entries, H is the *Walsh-Hadamard matrix* and D is a diagonal matrix of iid sign random variables, equal to ±1 with probability 1/2. Recall that the Walsh-Hadamard matrix is defined recursively by

$$H_n=\begin{pmatrix}H_{n/2} & H_{n/2}\\ H_{n/2} & -H_{n/2}\end{pmatrix},$$

with $H_1=(1)$. This requires n to be a power of two. For n that is not a power of two, we discuss more general constructions below.

For S defined in (III.25), we denote the action of the Walsh-Hadamard matrix H and the signflip matrix D on a vector w_i as

$$z_i:=\frac{1}{\sqrt{n}}HDw_i,\quad 1\leqslant i\leqslant k.$$

Note that each entry $z_i(l)$ is of the form

$$z_i(l)=\sum_{j=1}a_j^{(l)}w_i(j),$$

where $a_j^{(l)}=\pm n^{-1/2}$ is chosen independently and uniformly. Then a Chernoff type bound gives that the z vectors are delocalized, i.e.,

$$\|z_i\|_\infty\leqslant C\frac{\log n}{\sqrt{n}} \tag{III.26}$$

with high probability. Moreover, $\{z_i\}$ are orthonormal since $HD/\sqrt{n}$ is orthogonal. Then the result for uniform random

sampling can be applied here without the delocalization assumption in (III.24), because (III.26) already gives the desired delocalization for z_i-s after acting HD on w_i-s.

The argument above applies more broadly to general *Hadamard matrices*. An $n \times n$ possibly complex-valued matrix H is called a Hadamard matrix if $H/\sqrt{n}$ is orthogonal and the absolute values of its entries are unity, $|H_{ij}| = 1$ for $i, j = 1, \dots, n$. The Walsh-Hadamard matrix above clearly has these properties. Another construction is the discrete Fourier transform (DFT) matrix with the (u, v)-th entry equal to $H_{uv} = e^{-2\pi i(u-1)(v-1)/n}$. Multiplying this matrix from the right by X is equivalent to applying the discrete Fourier transform to each column of X, up to scaling. The time complexity for the matrix-matrix multiplication for both transforms is $\mathrm{O}(np \log n)$ using the Fast Fourier Transform.

To summarize, we have the following theorem as a corollary of Theorem III.3 and the delocalization property in (III.26).

Theorem III.4 (Randomized Hadamard sampling)**.** *Suppose that the assumptions in Theorem III.1 hold except that S is now a random sampling matrix as in* (III.25)*, where H is a general $n \times n$ Hadamard matrix. Then the results* (III.15)–(III.18) *hold.*

See Figure 6 for experimental results supporting this theorem.

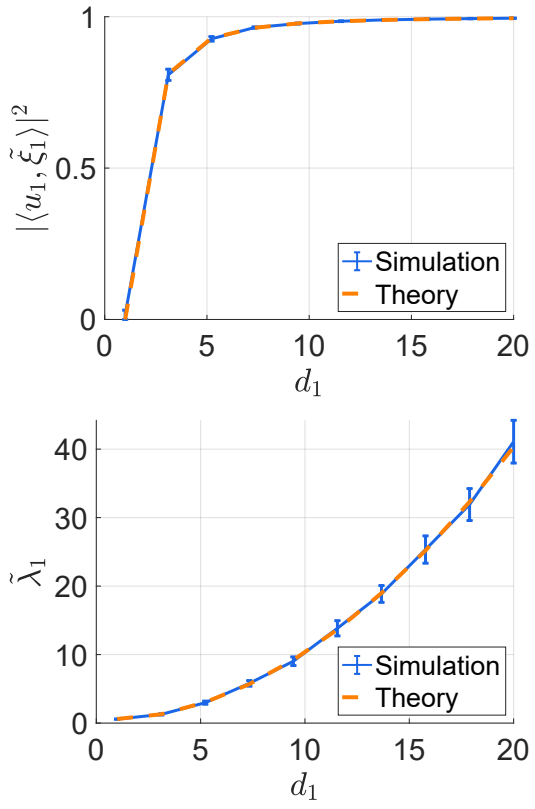

Fig. 6: Checking the accuracy of the spiked eigenvalue and eigenvector formulas for the subsampled randomized Hadamard transform. We follow the protocol from the experiment in Figure 3.

Gaussian data. From Section III-A1, we find the same results for Gaussian noise if we can show $W^\top S^\top S W \approx r/n \cdot I_k$. This follows from the same argument as for uniform sampling.

E. CountSketch

Another popular sketching method is CountSketch [18], also known as Clarkson-Woodruff sketch [22]. Here S is an $r \times n$ matrix that has a single randomly chosen non-zero entry $S_{h(j),j}$ in each column j, for a uniformly random mapping $h : \{1, \dots, n\} \to \{1, \dots, r\}$. Moreover, each $S_{h(j),j}$ is a Rademacher random variable, i.e., $S_{h(j),j} = \pm 1$ with probability 1/2. In other words, we have

$$S_{ij} = \delta_{ih(j)} a_j, \tag{III.27}$$

where a_j are iid Rademacher random variables that are independent of h. Intuitively, S maps the vector x to a random partition of its entries (mapping into random buckets), and takes randomly signed sums of the entries in each element of the partition (or in each bucket).

When applied to an $n \times p$ matrix X, SX computes an $r \times p$ matrix, such that each row is a randomly signed sum of some rows of X. This is similar to random sampling. However, the advantage is that no rows of X are "left out", and thus we automatically get a type of adaptive leverage score sampling, see e.g., [22]. The only constraint is that we need r to be large enough so that we avoid collisions of rows with large leverage scores.

In our case, it turns out that it is advantageous to study a slightly modified "normalized" CountSketch. To see this, we denote

$$SS^\top = \mathrm{diag}(c_1, \dots, c_r),$$

that is, c_i is the number of coordinates from $1, \dots, n$ that map into the i-th bucket. Then $(c_1, \dots, c_r)$ has the exact joint distribution

$$(c_1, \dots, c_r) \sim Multinomial(n; 1/r, \dots, 1/r).$$

Each c_i has a marginal distribution $\sim Binomial(n, 1/r)$, with mean n/r, and variance $\frac{n}{r}(1 - \frac{1}{r})$. As $r/n \to \xi > 0$ when $n \to \infty$, c_i converges in distribution to a Poisson random variable with constant rate. We know that the Poisson distribution is unbounded, which gives that for any fixed constant $C > 0$,

$$\limsup_{n \to \infty} \mathbb{P}(\|SS^\top\| \geqslant C) \geqslant c$$

for some constant $c > 0$ depending only on C and ξ. Hence *the operator norm of the sketching matrix is unbounded* (i.e., the first bound in (A.7) in our derivation fails) with non-zero probability. This is a problem because, theoretically, the spikes may be "covered up" by the noise eigenvalues. For example, there exists an event, say Ξ, of non-zero probability such that there is a small portion of c_i-s that are at least, say $C = 10^3$. Then on Ξ, the limiting noise singular value spectrum of SX is very wide, and hence can cover up the true signals. Such an event Ξ holds with small (but non-zero) probability, which is not a problem in applications if n/r is large. However, it prevents us from getting results that hold with probability $1 - \mathrm{o}(1)$ in the setting with $\limsup_{n \to \infty} n/r < \infty$. Moreover,

the spectral distribution of $SS^\top$ spreads out widely, but we will see that it is better to have a more concentrated spectral distribution (cf. Remark III.7). Hence we propose a simple normalization, in which we divide each bucket by the square root of the number of entries mapped into it. Formally, we define $\widehat{S} := (SS^\top)^{-1/2}S$, such that $\widehat{S}\widehat{S}^\top = I_r$. Then we shall use $\widehat{S}$ as our sketching matrix. With ξ_n converging to a constant, there is a significant number of zeros among the counts. Hence $(SS^\top)^{-1/2}$ should be understood as a pseudo-inverse. Alternatively, we can discard the buckets of size zero at the beginning.

Experiments show that the simple normalized version of CountSketch works similarly to uniform projection. As discussed above, we normalize SX as $B^{-1/2}SX$, where $B = SS^\top$ is the matrix of counts mapped into each bucket. In the regime where n/r is a constant, the probability of getting a zero count is approximately

$$\mathbb{P}(Poisson(1/\xi_n) = 0) = \exp(-1/\xi_n) = \exp(-n/r).$$

We discard those rows. From Figures 7, 8, 11 and 13, we find that the value of $|\langle u_i, \widetilde{\xi}_i\rangle|^2$ for normalized CountSketch is larger—if only slightly—than the one for CountSketch. This shows that normalized CountSketch is more accurate than the original CountSketch, in the sense that the principal components of the sketched matrix approximate the principal components of the signal matrix in a better way. The reason is that CountSketch has some large buckets, and the sum of the rows mapped into them can sometimes dominate the eigenvectors, leading to a loss of precision. In Remark III.7 below, we will also give another heuristic explanation. (Note that in Figure 7 in particular, normalized CountSketch and the original one have similar accuracy. This is because $r/n = 0.1$ is small and hence $SS^\top$ concentrates well around $\frac{n}{r}I_r$.)

In Figure 8, we compare the accuracy of CountSketch and normalized CountSketch where $p = 500$, $n \in \{20, 50, 100, 500\}$, and $\xi = 0.2$. We see that normalized CountSketch is more accurate than the unnormalized version, especially for small n and large p. However, the standard errors overlap, so one must exercise some caution when reading these figures. This simulation also shows that our theoretical formula is accurate even when n, p, and r are relatively small.

If $n \gg r$, say $n \geqslant Cr\log n$ for some large constant $C > 0$, then each c_i concentrates around n/r. In much of the literature on sketching, this is a common assumption [78], [104]. In this case, we have

$$\widehat{S} \approx \xi_n S,$$

and hence $\widehat{S}$ is simply a rescaling of the CountSketch matrix S, and thus immaterial.

We now state the theoretical results for CountSketch. The proof is presented in Section D.

Theorem III.5 (CountSketch)**.** *Suppose that the assumptions in Theorem III.1 hold, except we assume the delocalization condition* (III.24) *and that S is a random sampling matrix as in* (III.27)*. Then* (III.15)–(III.18) *hold if we replace ξ with $\widehat{\xi} = \xi\,[1 - \exp(-1/\xi)]$.*

CountSketch can be regarded as an interpolation between the uniform random sampling and randomized Hadamard sampling. For the time complexity, we have "uniform random sampling < CountSketch < randomized Hadamard sampling". However, unfiorm random sampling and CountSketch are much closer in complexity (within a constant), while randomized Hadamard sampling has an additional logarithmic factor

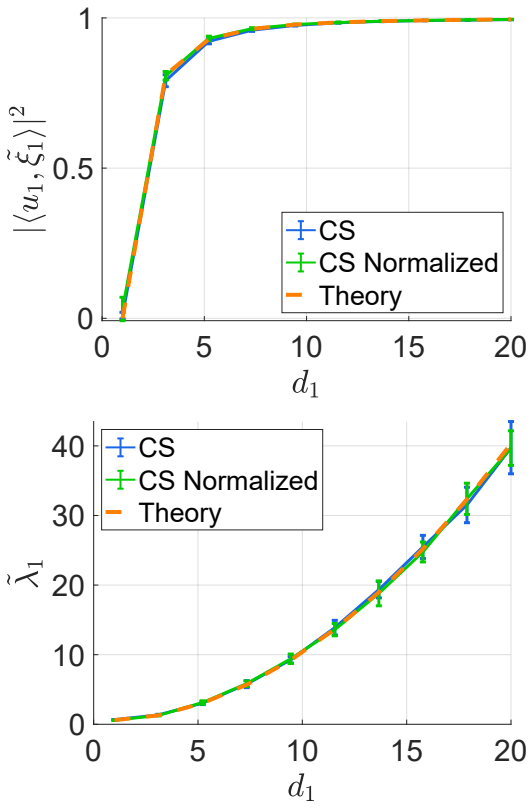

Fig. 7: Checking the accuracy of the spiked eigenvalue and eigenvector formulas for CountSketch. We follow the protocol from the experiment in Figure 3. Here we rescale $\tilde{\lambda}_1$ for CountSketch by a factor of r/n.

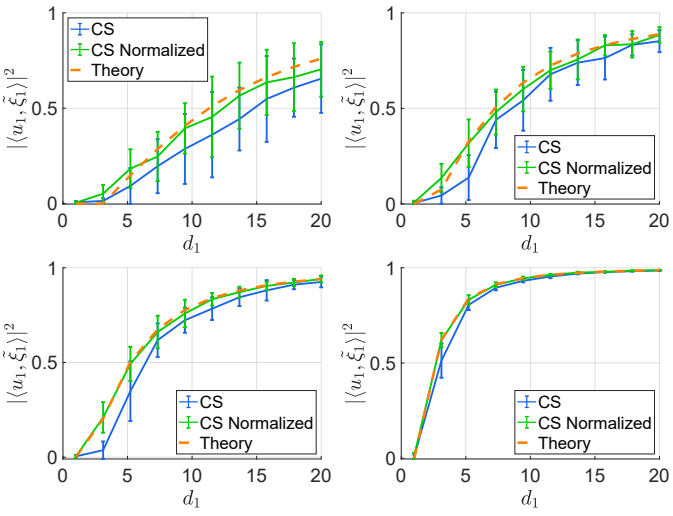

Fig. 8: Comparing the accuracy of CountSketch and normalized CountSketch. We take $n \in \{20, 50, 100, 500\}$, $p = 500$, and $r/n = 0.2$. The setting is the same as in Figure 3.

in the cost. On the other hand, uniform random sampling and the CountSketch requires the delocalization condition (III.24), while randomized Hadamard sampling does not.

One of the advantages of CountSketch is that it is extremely fast for sparse datasets. For example, we can consider sparse sample covariance matrices which are Hadamard products of the form $\widetilde{X} = A \circ X$, where X is a random matrix considered in this paper and A is a random matrix with iid $Beroulli(p_n)$ entries. Then $0 < p_n < 1$ controls the sparsity of the sample covariance matrices. We expect that CountSketch will perform well in the sparse case with $p_n \ll 1$. Unfortunately, this case is beyond our current setting—the moment condition (A.8) will be violated if the entries of $\widetilde{X}$ are scaled to have variance n^{-1}. However, we expect that our results still hold under the sparse setting, although we need to rebuild the whole theory in [106] from scratch using the methods in [47] for sparse Erdős-Rényi graphs. This is beyond the scope of the current paper, and we will explore this topic in future work.

We also remark that the delocalization condition (III.24) is needed for CountSketch because we are considering the setting where r is of the same order as n. In the conventional setting where $n \gg r \gg p$, this condition is not needed [22]. In our setting, we can recover this result. With a simple Chernoff inequality and a union bound, we know that if $n \geqslant Cr \log r$ for a large enough constant $C > 0$, then with probability $1 - \mathrm{o}(1)$ all the c_i-s are concentrated around n/r. Moreover, as for Hadamard sampling in Section III-D, $\widehat{S} w_i$ will be delocalized, i.e., $\|\widehat{S} w_i\|_\infty \to_P 0$ as $n \to \infty$. This bound holds for the same reason as (III.26), because we take random averages over roughly n/r entries of w_i.

F. Strong signals

Finally, in this subsection, we consider a more general spiked covariance matrix model

$$Y = \sum_{i=1}^{k} d_i w_i u_i^\top + X\Sigma^{1/2}, \tag{III.28}$$

where the covariance matrix Σ can be non-identity. In this case, the covariance matrix of the spiked model is of the form

$$\widetilde{\Sigma} = \Sigma + \sum_{i=1}^{k} d_i^2 u_i u_i^\top. \tag{III.29}$$

Since u_i's are not necessarily the eigenvectors of Σ, they are also not the eigenvectors of $\widetilde{\Sigma}$ in general. However, if we assume the signal strengths to be sufficiently large, then we can regard u_i as an approximate eigenvector of $\widetilde{\Sigma}$. This is the setting we shall consider in this subsection.

We write $x = \mathrm{O}(y)$ if $|x| \leqslant C|y|$ for some constant $C > 0$ that does not depend on n or l_i. We will denote $a \wedge b = \min(a, b)$ and

$$l_i := d_i^2 \wedge \min_{j \neq i} |d_i^2 - d_j^2|.$$

Combining the arguments in the proof of Theorem III.1 with standard perturbation theory for matrices, we can obtain the following theorem. The proof of will be given in Appendix E.

Theorem III.6 (Large signals)**.** *Suppose that the assumptions in Theorem III.1 hold, so we consider uniform orthogonal random projections. Moreover, assume that for some fixed $k_+ \leqslant k$,*

$$\max_{1 \leqslant i \leqslant k_+} l_i \geqslant C_0 \|\Sigma\| \tag{III.30}$$

for a sufficiently large constant $C_0 > 0$ that does not depend on n, p, Σ or l_i. Then for any $1 \leqslant i \leqslant k_+$, we have

$$\widetilde{\lambda}_i \to_P \theta_i = \xi(d_i^2 + E_{ii}) + \gamma\rho_1 + \mathrm{O}(l_i^{-1}). \tag{III.31}$$

Here $E := U^\top \Sigma U$, and ρ_i are the moments of the spectral distribution of Σ,

$$\rho_i := \int x^i \pi_\Sigma(\mathrm{d}x). \tag{III.32}$$

Also,

$$\begin{aligned} &|\langle u_i, \widetilde{\xi}_i \rangle|^2 \\ &\to_P \frac{\xi - \frac{\gamma}{d_i^4}\rho_2}{\xi + \frac{\gamma}{d_i^2}\left[\rho_1 + d_i^{-2}(\rho_2 - \rho_1 E_{ii})\right]} + \mathrm{O}(l_i^{-3}), \end{aligned} \tag{III.33}$$

and for $j \neq i$,

$$\begin{aligned} &|\langle u_j, \widetilde{\xi}_i \rangle|^2 \\ &\to_P \frac{\xi - \frac{\gamma}{d_i^4}\rho_2}{\xi + \frac{\gamma}{d_i^2}\left[\rho_1 + d_i^{-2}(\rho_2 - \rho_1 E_{ii})\right]} \left| \frac{E_{ji}}{d_i^2 - d_j^2} \right|^2 + \mathrm{O}(l_i^{-3}). \end{aligned}$$

Similarly, if the assumptions in Theorem III.3 (uniform random sampling) or Theorem III.4 (Hadamard transform) hold, then the same results hold; if the assumptions in Theorem III.5 (CountSketch) hold, then the same results hold if we replace ξ with $\widehat{\xi}$.

We check the formulas in simulations. In the first example, we take $\Sigma = O^\top \Lambda O$, where O is a $p \times p$ orthogonal matrix, Λ is a diagonal matrix with $\Lambda_{11} = 5$, $\Lambda_{ii} = 2$ for $2 \leqslant i \leqslant p/2$, $\Lambda_{ii} = 1$ for $p/2 < i \leqslant p$; see Figure 9. In the second example, we take Σ to be the Toeplitz matrix with $\Sigma_{ij} = 0.9^{|i-j|}$; see Figure 10. In general, for a Toeplitz matrix whose (i, j)-th entry is $q^{|i-j|}$, we have $\rho_1 = p^{-1}\operatorname{Tr}(\Sigma) = 1$ because the diagonal entries are all ones. For ρ_2, we have

$$\rho_2 = \frac{1}{p}\operatorname{Tr}(\Sigma^2) = \frac{1}{p}\|\Sigma\|_F^2 = \frac{1}{p}\sum_{i,j}\Sigma_{ij}^2.$$

Among the p^2 entries of Σ, the p diagonal entries are equal to 1; and for $i = 1, \ldots, p-1$, there are $2(p-i)$ entries that are equal to q^i. Thus,

$$\begin{aligned} \frac{1}{p}\sum_{i,j}\Sigma_{ij}^2 &= 1 + \frac{1}{p}\sum_{i=1}^{p-1} 2(p-i)q^{2i} \\ &= 1 + \frac{2}{p}\left[\frac{pq^2}{1-q^2} - \frac{q^2(1-q^{2p})}{(1-q^2)^2}\right], \end{aligned}$$

which converges to $\frac{1+q^2}{1-q^2}$ as p goes to infinity. We see a good match between the simulation and the theoretical result, as long as the signal strength d_i is reasonably large.

In fact, we can get more precise results by deriving higher order asymptotic expansions in terms of l_i^{-1}. The calculations become more tedious, so we do not pursue this direction here.

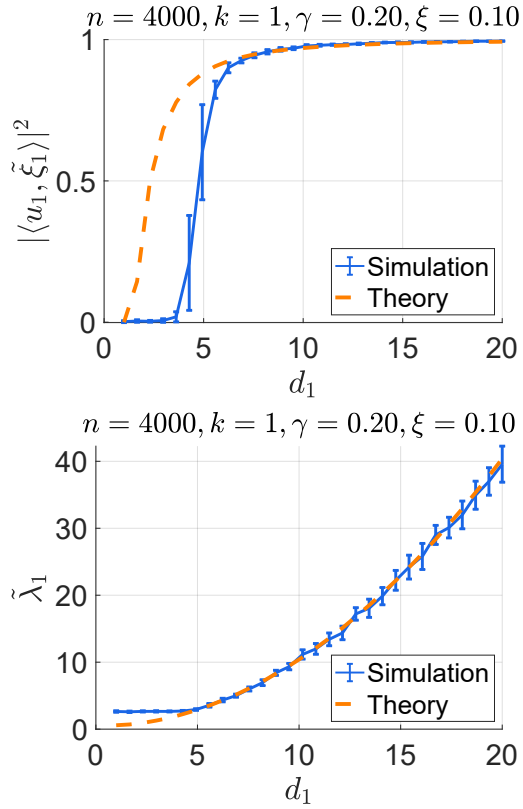

Fig. 9: Checking the accuracy of the spiked eigenvalue and eigenvector formulas for large signals in a single-spiked model. We follow the protocol from the experiment in Figure 3. Here $\Sigma = O^\top \Lambda O$, where O is a $p \times p$ orthogonal matrix, Λ is a diagonal matrix with $\Lambda_{11} = 5$, $\Lambda_{ii} = 2$ for $2 \leqslant i \leqslant p/2$, $\Lambda_{ii} = 1$ for $p/2 < i \leqslant p$, and d ranges from unity to 20 with equal spaces.

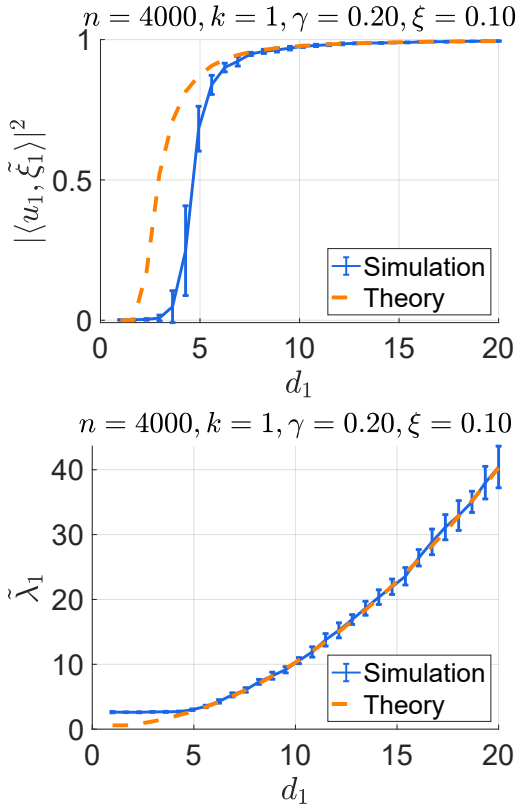

Fig. 10: The protocol is the same as in Figure 9, except that Σ is the Toeplitz matrix with the (i, j)-th entry equal to $0.9^{|i-j|}$.

Remark III.7. The two terms on the right-hand side of (III.31) can be understood heuristically as follows. First, based on standard perturbation theory, the i-th largest eigenvalue of $\widetilde{\Sigma}$ is approximately $d_i^2 + u_i^\top \Sigma u_i = d_i^2 + E_{ii}$. As discussed in Section II-A, heuristically after projection into r-dimensional subspace, the signal strength should go down by a factor of ξ, which leads to the term $\xi(d_i^2 + E_{ii})$ in (III.31). For the $\gamma\rho_1$ term, we consider the extreme case where $\xi \to 0$ and hence the signal strength goes down to zero. Without loss of generality, we assume that S is random sampling. Then by concentration of measure, one can see that

$$(SX\Sigma^{1/2})(\Sigma^{1/2}X^\top S^\top) \approx \left(\frac{1}{n}\sum_{i=1}^{p}\Sigma_{ii}\right) I_{r\times r} = \gamma\rho_1 I_{r\times r}.$$

This leads to the $\gamma\rho_1$ term that does not depend on ξ.

This analysis also allows to gain some insights into the comparison of different sketching methods. For simplicity, suppose U is a uniform partial orthonormal matrix. Then E_{ii} is well-concentrated around ρ_1. Now we notice that for a fixed ρ_1, the right-hand side of (III.33) becomes smaller as ρ_2 increases. In particular, the second moment of the spectral distribution of Σ is minimized when it is degenerate (i.e. concentrates on one point).

With a similar method, in the setting with $\Sigma = I_p$, a general sketching matrix S and a uniform partial orthonormal W, we can derive that

$$|\langle u_i, \tilde{\xi}_i \rangle|^2 \to_P \frac{\xi - \frac{\rho_2}{\rho_1^2}\frac{\gamma}{d_i^4}}{\xi + \frac{\rho_2}{\rho_1^2}\frac{\gamma}{d_i^2}} + \mathrm{O}(l_i^{-3}),$$

where ρ_1 and ρ_2 are the first and second moments of the spectral distribution of $SS^\top$. We omit the details, since the derivation is similar to the one in Appendix E. One can also compare it with (III.22). Hence it is better to use a sketched matrix with smaller ρ_2/ρ_1^2, which is minimized at unity when the spectral distribution of $SS^\top$ is degenerate. This heuristically explains why projections with iid entries and CountSketch are slightly worse than other methods—they have less concentrated spectrum compared to other methods.

To our knowledge, a model at the level of generality of (III.28) has not been studied in the literature, even in the strong signal regime. In the classical setting, it is usually assumed that Σ is identity or a finite rank perturbation of identity matrix; see e.g., [6], [7], [8], [10], [29], [66] etc. This is also our setting in Sections III-A to III-E. Another type of spiked covariance model has spikes added to the population covariance matrices directly; see e.g., [9], [13], [30], [84] etc. That model is given by $X\widehat{\Sigma}^{1/2}$, where $\widehat{\Sigma}$ is a spiked covariance matrix of the form (III.29). By diagonalizing the matrix $\widehat{\Sigma}$, one can assume that u_i-s are also eigenvectors of $\widehat{\Sigma}$, which is more restrictive than our model (III.28). Thus we believe that the Theorem III.6 and

the method used in its proof may be of independent theoretical interest.

IV. Empirical support

A. Proposed method

We aim to verify our results empirically. In previous work for linear regression [35] we developed formulas for the behavior of the OLS residuals under sketching. We predicted the behavior of the ratio of residuals, as a function of the known quantities n, p, r only. This idea is similar to constructing a pivotal random variable in statistical inference, whose behavior does not depend on un-measured quantities. Surprisingly, we found that the ratio of residuals can be close to the predicted value in empirical datasets.

Here we do not have the direct analogue of the residuals. However, we can work from first principles to derive a similar method. We know that the top eigenvalues in standard spiked models follow the spiked forward map $\ell \to \lambda(\ell, \gamma)$ from equation (III.1) (see e.g., [6], [7]). A well known method to estimate the spike is to invert this map. These methods have been implemented in the `EigenEdge` package [32].

Thus, we propose to calculate the inverse both for the original and sketched data. In our model, both should be close to ℓ. We propose the statistic

$$T = \frac{\lambda^{-1}\left(\sigma_1(X)^2, p/n\right)}{\lambda^{-1}\left(\sigma_1(SX)^2, p/r\right)}.$$

Our theoretical results predict that we should have $T \approx 1$.

B. Datasets tested

We consider three data sets to test our theoretical results: the Human Genome Diversity Project (HGDP) dataset (e.g., [16], [72]), the Million Song Dataset (MSD, [11]) and New York Flight Dataset [103]. For each, we take uniform orthogonal random projections on the data with $r = \lfloor \xi n \rfloor$, with $\xi = 0.8, 0.5, 0.3$. For HGDP, we repeat this while subsampling (1) every 20th column; (2) every 10th row and 20th column. These values of ξ are relatively high, compared to values that practitioners may be interested in for large computational savings. The reason is that the current asymptotic framework, with a fixed number of spikes just above the noise level, is expected to be accurate when r is not too small. Providing accurate predictions of the behavior of the spikes in the regime where $r/n = o(1)$ is not covered by our current work and remains important future work.

For some context, the purpose of collecting the HGDP dataset was to evaluate the diversity in the patterns of genetic variation across the globe. We use the CEPH panel, in which single nucleotide polymorphism (SNP) data was collected for 1043 samples representing 51 different populations from Africa, Europe, Asia, Oceania and the Americas. We obtained the data from www.hagsc.org/hgdp/data/hgdp.zip. We provide the data and processing pipeline on this paper's GitHub page.

The data has $n = 1043$ samples, and we focus on the $p = 9730$ SNPs on chromosome 22. Thus we have an $n \times p$ data matrix X, where $x_{ij} \in \{0, 1, 2\}$ is the number of copies of the minor allele of SNP j in the genome of individual i. We standardize the data SNP-wise, centering each SNP by its mean, and dividing by its standard error. For this step, we ignore missing values. Then, we impute the missing values as zeroes, which are also equal to the mean of each SNP.

For the HGDP dataset, we have seen in previous work that it is not well modeled by a matrix with iid Gaussian entries [36]. In particular, there are correlations both between the columns and between the rows. Despite this model mismatch, for $\xi = 0.8$ we find values of T between 1.2 and 1.4 on this dataset, which are quite close to the expected value of unity under correct model specification. This suggests that our theory may sometimes be applicable and relevant even when the data do not follow the theoretical model.

We also consider the Million Song Dataset [11] and New York Flights Dataset [103]. For $\xi = 0.8$, we find $T = 1.26$ and $T = 1.24$, respectively. However, for both datasets, if we use $r = 0.5n$, then T is about 2; if $r = 0.3n$, T is about 3. This also suggests that, in these datasets where the assumptions does not hold, the theoretical results become somewhat less accurate as r/n decreases.

C. Simulation for multi-spike model

From Section III-A to Section III-F, the simulations all concern the single-spiked model with rank $k = 1$. Here we verify our theoretical results on a multi-spiked model with $k = 5$. The results for $|\langle u_i, \widetilde{\xi}_i \rangle|^2$ and $\widetilde{\lambda}_i$ are shown in Figures 11 and 12 respectively. We see that the formulas are very accurate for $\widetilde{\lambda}_i$, but less accurate for $|\langle u_i, \widetilde{\xi}_i \rangle|^2$ for large signals. Heuristically, this is because the variance of $|\langle u_i, \widetilde{\xi}_i \rangle|^2$ increases compared to the single-spike case due to the repulsion between different spikes, see e.g. [84].

V. Discussion

We have chosen to present results with convergence in probability in this paper, because we want our assumptions to be as general as possible. However, with certain stronger assumptions, it is possible to improve the results to almost sure convergence.

One obstacle for improving the results is the convergence of the block $W^\top S^\top (1 + m_{1c}(x) S S^\top)^{-1} S W$ of the master matrix $M(x)$ in equation (II.3). In the manuscript, we only show convergence in probability. To have a stronger convergence for uniform random sampling and CountSketch, we need conditions on the vectors w_i that are stronger than (III.24). For uniform random sampling, suppose we assume a stronger delocalization condition on the vectors w_i,

$$\max_{1 \leqslant i \leqslant k} \|w_i\|_\infty \leqslant c(\log n)^{-1}$$

for a small enough constant $c > 0$. Then with a Chernoff or Bernstein type matrix concentration bound, we can obtain that

$$\left\| W^\top S^\top S W - \frac{r}{n} I_k \right\| = \mathrm{O}((\log n)^{-1/2}),$$

with probability $1 - \mathrm{O}(e^{-C \log n})$ for a large constant $C > 0$. Then using the Borel-Cantelli lemma, we can show that $W^\top S^\top S W$ converges almost surely to $\frac{r}{n} I_k$.

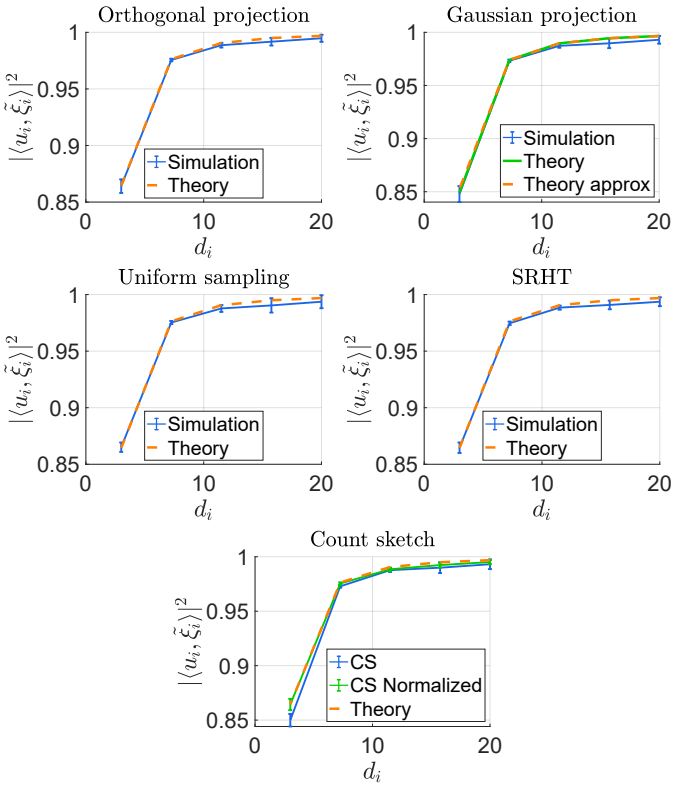

Fig. 11: Checking the accuracy of the formulas for $|\langle u_i, \widetilde{\xi}_i\rangle|^2$ for different sketching methods. We take $n = 20000$, $p = 2500$, $r = 2000$, $k = 5$. We plot the mean and standard error over 20 repeated experiments. The signal strengths $d_1, \ldots, d_5$ range from 20 to 3 with equal spaces.

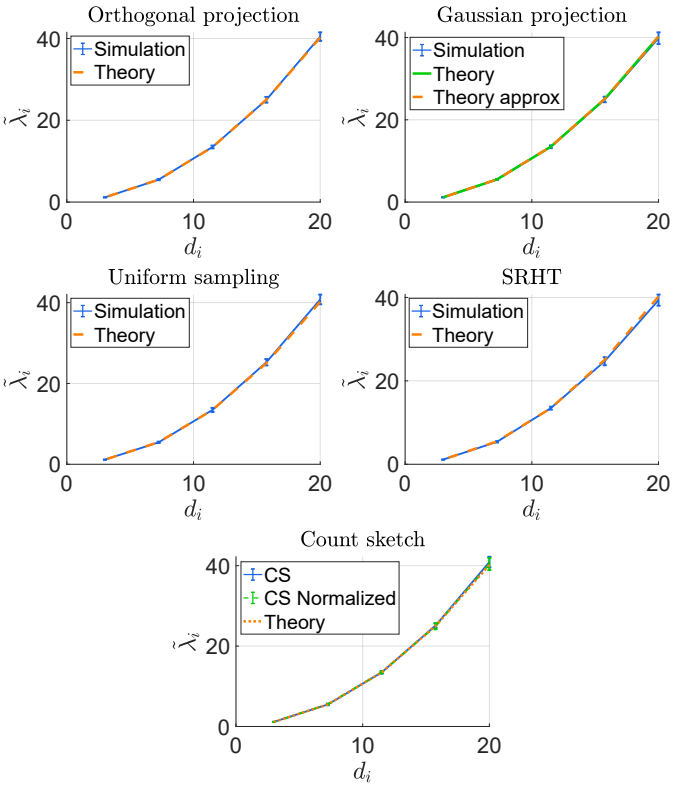

Fig. 12: Checking the accuracy of the formulas for $\widetilde{\lambda}_i$ for different sketching methods. The protocol is the same as Figure 11.

Another obstacle is the moment assumption on the matrix entries of X. Right now, we assume that the entries of X have a finite $(4+\varepsilon)$-th moment as in equation (A.8). Then we can truncate the entries as in (F.1) such that for a small constant δ, $\max_{i,j}|x_{ij}| \leqslant n^{-\delta}$ on an event with probability $1 - \mathrm{O}(n^{-\delta})$. The bounded entry condition $\max_{i,j}|x_{ij}| \leqslant n^{-\delta}$ is necessary for our Theorems A.7 and A.8, but the probability $1-\mathrm{O}(n^{-\delta})$ is not sufficient. To improve this probability, we need a stronger moment assumption. For example, if the entries of X have finite $(6+\varepsilon)$-moment, then we can truncate the entries of X such that for a small constant δ, $\max_{i,j}|x_{ij}| \leqslant n^{-\delta}$ on an event with probability $1 - \mathrm{O}(n^{-1-\delta})$. Then we can use the Borel-Cantelli lemma to improve the results to almost sure convergence.

In addition, in future work, it may be of interest to investigate other sketching methods that have been proposed. In particular, uniform sampling can work poorly when the data are highly non-uniform, because some datapoints are more influential than others for the PCs. There are more advanced sampling methods that sample each row of X with some non-uniform probability π_i which relates to the importance of the ith sample, such as d_2 sampling [41], where π_i is proportional to the squared norm of the ith row, or leverage score sampling, where the scores are proportional to the leverage scores [21], [77], [78].

Another frequently used type of random projections are the so-called oblivious sparse norm-approximating projections (OSNAPs) [69], [82]. More precisely, an $r \times n$ random projection matrix S is an OSNAP if $S_{ij} = \delta_{ij}\sigma_{ij}/\sqrt{s}$, where $s \geqslant 1$ is a fixed integer, σ_{ij} are random signs, and δ_{ij} are indicator random variables satisfying the following properties:

- fixed number of nonzeros per column: for any $1 \leqslant j \leqslant n$, $\sum_{i=1}^{r}\delta_{ij} = s$ with probability 1;
- negative correlation between the nonzeros: for any $E \subset \{1, \cdots, r\} \times \{1, \cdots, n\}$, $\mathbb{E}\prod_{(i,j)\in E}\delta_{ij} \leqslant (s/r)^{|E|}$.

A concrete example is when we independently choose s nonzero locations for each column, uniformly at random over all possible subsets of size s.

The difficulty in analyzing leverage score sampling and OSNAP lies in a complete understanding of the exact ESD of $SS^\top$, which is needed in both the study of the self-consistent equations in (II.2) and the matrix (II.3). However, if the signals are strong, then it is possible to obtain some approximate results using the argument in the proof of Theorem III.6, where only the first few moments of the ESD of $SS^\top$ is needed.

We can compare leverage score sampling and OSNAP with the sketching methods analyzed in Section III through simulations; see Figure 13. We take $n = 4000$, $p = 800$, $k = 8$ with varying signal strengths and different r. The error bars are the standard deviations over 20 independent repetitions. We plot the overlap between the spiked population eigenvector and the sample eigenvector after sketching. From Figure 13,

we observe the following common phenomena:

- Haar projection, uniform sampling, subsampled randomized Hadamard transform, and normalized CountSketch have roughly the same efficiency, and they are all better than iid Gaussian projection, as discussed in Section III-B. Moreover, unnormalized CountSketch is less accurate than normalized CountSketch. (However, of course, CountSketch can have other advantages like running time adapted to input sparsity.)
- Leverage score sampling behaves similarly to uniform sampling. Note that this is related to the choice of model in this paper. When the data is highly non-uniform, we expect that leverage score sampling will be better than uniform sampling.
- OSNAP is less accurate than all the other methods. Again, this method can have other advantages, like near-optimally small r to ensure oblivious subspace embedding for sparse inputs. Moreover, the gap between OSNAP and other sketching methods gets smaller as r increases.

For Haar projection, uniform sampling, subsampled randomized Hadamard transform, and normalized CountSketch, the ESD of $SS^\top$ is a singleton at unity. On the other hand, for iid Gaussian projection, unnormalized CountSketch and OSNAP, the ESD of $SS^\top$ is supported on an interval around unity. Thus based on the simulations and the discussion in Remark III.7, we see that in order to better preserve the eigenspace of the signal, it is better to have a more "concentrated" ESD for $SS^\top$.

Finally, it could be of interest to generalize the argument to methods designed for the streaming data setting, such as core sketching [99], and iterative methods that can achieve arbitrary accuracy, such as blanczos [62], [90] and randomized block Krylov iteration [80].

APPENDIX A
THE SKETCHED SPIKED MODEL

A. The model

We start by recalling the "low-rank-signal plus noise" or "spiked covariance" matrix model studied in the paper, with additional details to follow. The data is generated as

$$Y = \sum_{i=1}^{k} d_i w_i u_i^\top + X\Sigma^{1/2}.$$

Here $\sum_{i=1}^{k} d_i w_i u_i^\top$ is the signal component, $\{d_i\}_{1\leqslant i\leqslant k}$ give the strengths of the signals, and $\{w_i\}_{1\leqslant i\leqslant k}$ and $\{u_i\}_{1\leqslant i\leqslant k}$ are the left and right singular vectors of the signals, respectively. Also, $X\Sigma^{1/2}$ is the noise component, where Σ is a $p\times p$ deterministic covariance matrix, and $X=(x_{ij})$ is an $n\times p$ random matrix, where the entries x_{ij}, $1\leqslant i\leqslant n$, $1\leqslant j\leqslant p$, are real independent random variables satisfying

$$\mathbb{E}x_{ij}=0, \qquad \mathbb{E}|x_{ij}|^2 = n^{-1}. \tag{A.1}$$

We assume that the signal is independent of the noise matrix X. Such signal plus noise or spiked models have been widely studied. The special case $\Sigma = I_p$ is known as the *standard*

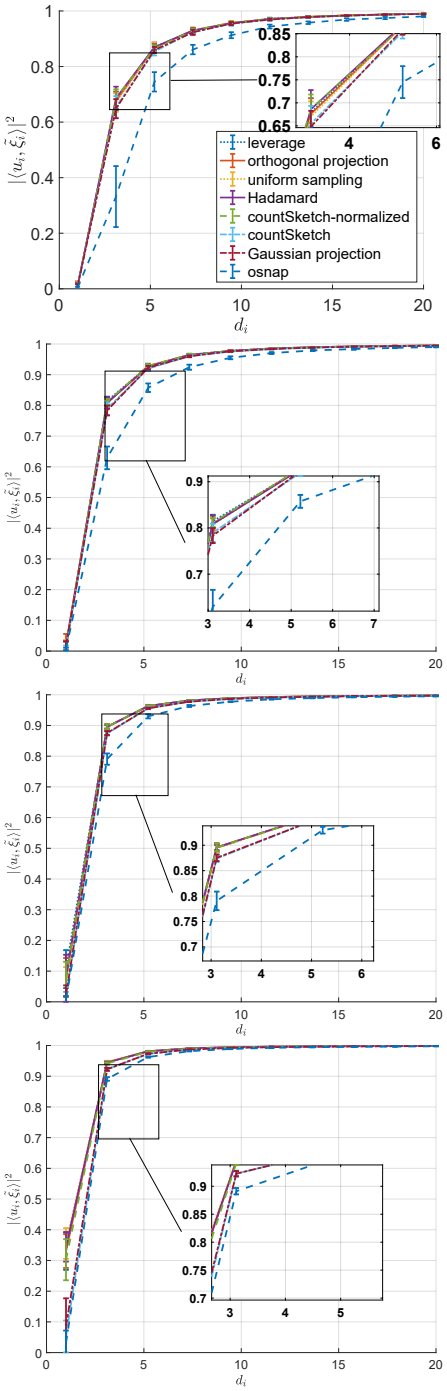

Fig. 13: Comparing different sketching methods. Here $n = 4000$, $p = 800$, $k = 1$, $r \in \{200, 400, 800, 1600\}$, and $x_{ij} \overset{iid}{\sim} \mathrm{Unif}(-\sqrt{3/n}, \sqrt{3/n})$. The error bars are the standard deviations over 20 independent repetitions. We order all methods according to their accuracy. The signal strengths d_1 range from unity to 20.

or (*Johnstone's*) spiked model [66], and more general spiked models have been proposed and studied, see [24], [85], [107] and references therein. When Σ is diagonal, and when w_i have iid entries, this model can be viewed as a specific *factor model*, and thus has a long history see e.g., [3], [34], [94], [97]. This model is fundamental for understanding principal component analysis (PCA), and has been thoroughly studied under high-dimensional asymptotics. Its understanding will serve as a baseline in our study.

In this paper, we are interested in the PCA of the *sketched* data matrix

$$\widetilde{Y} = SY$$

where S is an $r \times n$ random *sketching matrix* that is independent of both the signal and the noise. This can be written as

$$\widetilde{Y} = SX\Sigma^{1/2} + \sum_{i=1}^{k} d_i v_i u_i^\top, \quad v_i := Sw_i. \tag{A.2}$$

A similar spiked separable model has been studied in [30], although their setting is different, because the spikes are added to the population covariance matrices. However, we will still follow the presentation from [30] to some extent. We will study the spiked eigenvalues and eigenvectors of

$$\widetilde{\mathcal{Q}}_1 := \widetilde{Y}^\top \widetilde{Y} \in \mathbb{R}^{p\times p}, \quad \widetilde{\mathcal{Q}}_2 := \widetilde{Y}\widetilde{Y}^\top \in \mathbb{R}^{r\times r}.$$

We denote their (nontrivial) eigenvalues in descending order as $\widetilde{\lambda}_1 \geqslant \ldots \geqslant \widetilde{\lambda}_{p\wedge r}$. On the other hand, we will also use the non-spiked matrix

$$\widetilde{X} = SX\Sigma^{1/2}. \tag{A.3}$$

We denote the corresponding non-spiked matrices as

$$\mathcal{Q}_1 := \widetilde{X}^\top \widetilde{X} \in \mathbb{R}^{p\times p}, \quad \mathcal{Q}_2 := \widetilde{X}\widetilde{X}^\top \in \mathbb{R}^{r\times r},$$

with eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_{p\wedge r}$. We can study centered sample covariance matrices using our approach by setting

$$\widetilde{Y} = (I - ee^\top)SX\Sigma^{1/2} + (I - ee^\top)\sum_{i=1}^{k} d_i v_i u_i^\top,$$

with $e := r^{-1/2}(1, \cdots, 1)^\top \in \mathbb{R}^r$, or

$$\widetilde{Y} := S(I - ee^\top)X\Sigma^{1/2} + S\sum_{i=1}^{k} d_i (I - ee^\top) w_i u_i^\top,$$

with $e := n^{-1/2}(1, \cdots, 1)^\top \in \mathbb{R}^n$. In Appendix G, we show that this does not affect our results.

We assume that the number of signals k is a finite fixed integer, the strengths $d_1 > d_2 > \cdots > d_k > 0$ are fixed constants, and u_i, w_i are deterministic unit vectors. We shall consider the *high-dimensional* setting in this paper. More precisely, we assume that the aspect ratios

$$\gamma_n := p/n \to \gamma, \quad \xi_n := r/n \to \xi, \quad \text{as} \quad n \to \infty, \tag{A.4}$$

for some constants $\gamma \in (0, \infty)$ and $\xi \in (0, 1)$.

We assume that the noise covariance Σ and the outer product of the sketching matrix $B := SS^\top$ (an $r \times r$ matrix) have eigendecompositions

$$\Sigma = O_1 \Sigma_1 O_1^\top, \quad B = O_2 \Sigma_2 O_2^\top, \tag{A.5}$$

where

$$\Sigma_1 = \operatorname{diag}(\sigma_1, \cdots, \sigma_p), \quad \Sigma_2 = \operatorname{diag}(s_1, \cdots, s_r).$$

The eigenvalues of Σ and B are arranged in descending order as

$$\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_p \geqslant 0, \quad s_1 \geqslant s_2 \geqslant \ldots \geqslant s_r \geqslant 0.$$

We denote the empirical spectral distributions (ESDs) of Σ and $B = SS^\top$ by

$$\pi_\Sigma := \frac{1}{p}\sum_{i=1}^{p} \delta_{\sigma_i}, \quad \pi_B := \frac{1}{r}\sum_{i=1}^{r} \delta_{s_i}. \tag{A.6}$$

We assume that there exists a small constant $0 < \tau < 1$ such that for all n large enough,

$$\begin{aligned} &\max\{\sigma_1, s_1\} \leqslant \tau^{-1}, \\ &\max\{\pi_\Sigma([0,\tau]), \pi_B([0,\tau])\} \leqslant 1 - \tau. \end{aligned} \tag{A.7}$$

Both of these conditions are natural: the first condition means that the operator norms of Σ and B are bounded by τ^{-1}, and the second condition means that the spectra of Σ and B do not concentrate at zero. Moreover, we assume that π_Σ and π_B converge to certain probability distributions as $n \to \infty$. We will also assume some regularity conditions on π_Σ and π_B later.

Finally we assume that the random variables x_{ij} have finite (4+τ)-moments, in the following sense: there exists a constant $\tau > 0$ such that

$$\max_{i,j} \mathbb{E}|\sqrt{n}x_{ij}|^{4+\tau} \leqslant \tau^{-1}. \tag{A.8}$$

All constants appearing in our proof below may depend on the constants γ, ξ in (A.4) and the constant τ in (A.7), (A.8) and (A.16) below. However, for simplicity of presentation, we often do not indicate this dependence. Whenever we say "fix any constant C", this constant C does not depend on other constants (including γ, ξ and τ).

B. Resolvents and limiting laws

As usual in random matrix theory dating back to the seminal work of [79], we study the eigenvalue statistics of $\mathcal{Q}_{1,2}$ and $\widetilde{\mathcal{Q}}_{1,2}$ through their *resolvents* (or *Green's functions*). For quantities M_1, M_2 indexed by $l = 1, 2$, we will sometimes abbreviate M_1, M_2 by $M_{1,2}$. Throughout the following, we shall denote the upper half complex plane and the right half real line by

$$\mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}, \quad \mathbb{R}_+ := [0, \infty).$$

Definition A.1 (Resolvents)**.** *For $z = E + \mathrm{i}\eta \in \mathbb{C}_+$, and $l = 1, 2$ we define the following resolvents:*

$$\begin{aligned} \mathcal{G}_l(X, z) &:= (\mathcal{Q}_l(X) - z)^{-1}, \\ \widetilde{\mathcal{G}}_l(X, z) &:= (\widetilde{\mathcal{Q}}_l(X) - z)^{-1}. \end{aligned} \tag{A.9}$$

Note that the subscript "1" appears on $p \times p$ matrices, while the subscript "2" appears on $r \times r$ matrices. We denote the ESD $\rho^{(p)}$ of $\mathcal{Q}_1$ and its Stieltjes transform as

$$\begin{aligned} \rho \equiv \rho^{(p)} &:= \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\mathcal{Q}_1)}, \\ m(z) \equiv m^{(n)}(z) &:= \int \frac{1}{x - z}\rho^{(p)}(dx) = \frac{1}{p}\operatorname{Tr}\mathcal{G}_1(z). \end{aligned} \tag{A.10}$$

We also introduce the following quantities, which can be viewed as weighted Stieltjes transforms of $\mathcal{Q}_1$ and $\mathcal{Q}_2$, respectively:

$$\begin{aligned} m_1(z) \equiv m_1^{(n)}(z) &:= \frac{1}{n}\operatorname{Tr}\left(\Sigma \mathcal{G}_1(z)\right), \\ m_2(z) \equiv m_2^{(n)}(z) &:= \frac{1}{n}\operatorname{Tr}\left(B \mathcal{G}_2(z)\right). \end{aligned} \tag{A.11}$$

Notice that in equation (C.2) below, different from (A.11), we used the factor p^{-1} instead of n^{-1} in the definition of m_1^S. We adopted that notation there because $m_1^S(z)$ in that case actually plays the role of $m(z)$ in (A.10).

We now describe the limiting behavior of the density ρ and its Stieltjes transform $m(z)$. We consider the system of *self-consistent equations* for $(m_{1c}(z), m_{2c}(z)) \in \mathbb{C}_+^2$

$$\begin{aligned} m_{1c} &= \frac{1}{n}\sum_{i=1}^{p} \frac{\sigma_i}{-z(1+\sigma_i m_{2c})} \\ &= \gamma_n \int \frac{x}{-z\left[1+x m_{2c}(z)\right]} \pi_\Sigma(\mathrm{d}x), \\ m_{2c} &= \frac{1}{n}\sum_{\mu=1}^{r} \frac{s_\mu}{-z(1+s_\mu m_{1c})} \\ &= \xi_n \int \frac{x}{-z\left[1+x m_{1c}(z)\right]} \pi_B(\mathrm{d}x). \end{aligned} \tag{A.12}$$

It is known that this system admits a unique solution, see e.g., [43], [108]. Then we define m_c in terms of m_{2c}:

$$\begin{aligned} m_c(z) &:= \frac{1}{p}\sum_{i=1}^{p} \frac{1}{-z(1+\sigma_i m_{2c})} \\ &= \int \frac{1}{-z\left[1+x m_{2c}(z)\right]} \pi_\Sigma(\mathrm{d}x). \end{aligned} \tag{A.13}$$

It is easy to verify that $m_c(z) \in \mathbb{C}_+$ for $z \in \mathbb{C}_+$. It turns out that this is *the Stieltjes transform of the limiting spectral distribution* of the non-spiked sketched matrix $\mathcal{Q}_1 = \widetilde{X}^\top \widetilde{X}$, where $\widetilde{X}$ is defined in (A.3). We can recover the distribution of eigenvalues in the usual way, by inverting the Stieltjes transform. Letting $\eta \downarrow 0$, we obtain the probability measure ρ_c which describes the limiting distribution of the eigenvalues with the Stieltjes transform inversion formula

$$\rho_c(E) = \lim_{\eta \downarrow 0} \frac{1}{\pi} \operatorname{Im} m_c(E + \mathrm{i}\eta). \tag{A.14}$$

Moreover, under the assumption (A.7), the supremum of the support of $\rho_c(E)$ is at a finite value λ_+, known as "the right edge", which is also the "classical location" and almost sure limit of the largest eigenvalue of $\mathcal{Q}_1$. These known results are collected in the following lemma.

Lemma A.2 (Existence, uniqueness, and continuous density)**.** *For any $z \in \mathbb{C}_+$, there exists a unique solution $(m_{1c}, m_{2c}) \in \mathbb{C}_+^2$ to the systems of equations in (A.12), such that both functions m_{1c}, m_{2c} are Stieltjes transforms of two measures (not necessarily probability measures) μ_{1c} and μ_{2c} supported on $\mathbb{R}_+$. The function m_c in (A.13) is the Stieltjes transform of a probability measure μ_c supported on $\mathbb{R}_+$. Moreover, μ_c (resp. μ_{lc}, $l = 1, 2$) has a continuous density $\rho_c(x)$ (resp. $\rho_{lc}(x)$) on $(0, \infty)$, which is defined by* (A.14)*. The densities ρ_c and ρ_{lc} all have the same support on $(0, \infty)$, which is a union of intervals:*

$$\begin{aligned} \operatorname{supp} \rho_{1,2c} \cap (0,\infty) &= \operatorname{supp} \rho_c \cap (0,\infty) \\ &= \bigcup_{k=1}^{\mathfrak{a}} [e_{2k}, e_{2k-1}] \cap (0,\infty), \end{aligned} \tag{A.15}$$

where the number of components $\mathfrak{a} \in \mathbb{N}$ depends only on π_Σ and π_B. Here we order the components so that $e_{2k} < e_{2k-1} < e_{2k-2}$, hence e_1 is the supremum of the support–or "right edge"–of the densities. Under the first assumption in (A.7)*, we have $e_1 \leqslant C_\tau$ for a constant $C_\tau > 0$ depending on τ. Moreover, $m_{1c}(e_1) \in (-s_1^{-1}, 0)$ and $m_{2c}(e_1) \in (-\sigma_1^{-1}, 0)$.*

Proof. The proof of this lemma is contained in [108, Theorem 1.2.1], [60, Theorem 2.4] and [26, Section 3]. □

We shall call e_k the *spectral edges*. In particular, we will only focus on the rightmost edge $\lambda_+ := e_1$. Now we make the following assumption, which guarantees a classical square-root behavior of the spectral densities ρ_c near λ_+ and rules out the existence of spikes for $\mathcal{Q}_{1,2}$. In other words, the spikes of $\widetilde{\mathcal{Q}}_{1,2}$ are only caused by the signals in (A.2). We note that this is a mild condition, and holds in particular when the ESDs of Σ and B are well behaved. Specifically, when $B = I_n$ (i.e., when there is no projection), then it is known that the square root behavior holds as long as the limit of the ESD of Σ is sufficiently "regular" at its right edge. For instance, it is enough if the right edge of the limiting ESD is a point mass, or has a density bounded away from zero and infinity, see e.g., [5], [93]. This is a mild condition that, while possibly hard to check in applications, does not appear to be a significant limitation.

Assumption A.3 (Right edge regularity)**.** *There exists a constant $\tau > 0$ such that*

$$1 + m_{1c}(\lambda_+) s_1 \geqslant \tau, \quad 1 + m_{2c}(\lambda_+)\sigma_1 \geqslant \tau. \tag{A.16}$$

Under this assumption, we have the following lemma.

Lemma A.4 (Lemma 2.6 of [106], square root density at edge)**.** *Under assumptions* (A.4)*,* (A.7) *and* (A.16)*, there exists a value $a > 0$ of order 1 such that*

$$\rho_c(\lambda_+ - x) = a x^{1/2} + \mathrm{O}(x), \quad \text{as} \quad x \downarrow 0, \tag{A.17}$$

and

$$m_c(z) = m_c(\lambda_+) + \pi a (z - \lambda_+)^{1/2} + \mathrm{O}(|z - \lambda_+|), \tag{A.18}$$

as $z \to \lambda_+$. The bound (A.18) *also holds for $m_{1,2c}$ with possibly different constants $a_{1,2} > 0$.*

We introduce a classical self-adjoint linearization trick, dating back at least to Girko, see e.g., the works [57], [58], [59] and references therein. Define the *linearization* matrix as the following $(p+r)\times(p+r)$ self-adjoint block matrix, which is a linear function of X:

$$H \equiv H(X, z) := z^{1/2} \begin{pmatrix} 0 & \widetilde{X}^\top \\ \widetilde{X} & 0 \end{pmatrix}, \quad z \in \mathbb{C}_+, \tag{A.19}$$

where recall that $\widetilde{X} = S X \Sigma^{1/2}$ is the projected non-spiked matrix, and $z^{1/2}$ is taken to be the branch cut with positive

imaginary part. Then we define its resolvent (Green's function) as

$$G \equiv G(X,z) := (H(X,z)-z)^{-1}. \tag{A.20}$$

By the Schur complement formula and (A.9)

$$\begin{aligned} G(z) &= \begin{pmatrix} \mathcal{G}_1 & z^{-1/2}\mathcal{G}_1\widetilde{X}^\top \\ z^{-1/2}\widetilde{X}\mathcal{G}_1 & \mathcal{G}_2 \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{G}_1 & z^{-1/2}\widetilde{X}^\top\mathcal{G}_2 \\ z^{-1/2}\mathcal{G}_2\widetilde{X} & \mathcal{G}_2 \end{pmatrix}. \end{aligned} \tag{A.21}$$

Thus, an analysis of G yields directly an analysis of the resolvents $\mathcal{G}_{1,2}$. Similarly, we can define $\widetilde{H}$ and $\widetilde{G}$ by replacing $\widetilde{X}$ with the spiked version $\widetilde{Y}$. For simplicity of notation, we will sometimes use the index sets

$$\mathcal{I}_1 := \{1,...,p\}, \quad \mathcal{I}_2 := \{p+1,...,p+r\}, \quad \mathcal{I} := \mathcal{I}_1 \cup \mathcal{I}_2,$$
$$\mathcal{I}_2^n := \{p+1,\cdots,p+n\}$$

to label the indices of the matrices. For instance, since X is an $n\times p$ matrix, we will label its row indices according to $\mathcal{I}_2^n$, and its column indices according to $\mathcal{I}_1$:

$$X = (X_{\mu i})_{\mu\in\mathcal{I}_2^n, i\in\mathcal{I}_1},\ \Sigma = (\Sigma_{ij})_{i,j\in\mathcal{I}_1},\ S = (S_{\mu\nu})_{\mu\in\mathcal{I}_2,\nu\in\mathcal{I}_2^n}.$$

In the rest of this paper, we will consistently use the elements of the latin alphabet $i,j\in\mathcal{I}_1$ and the elements of the greek alphabet $\mu,\nu\in\mathcal{I}_2$ or $\mathcal{I}_2^n$.

We define the following matrix, which turns out to be the deterministic limit of the resolvent G of the linearization matrix H, as

$$\Pi(z) := \begin{pmatrix} \Pi_1 & 0 \\ 0 & \Pi_2 \end{pmatrix}, \tag{A.22}$$

where (using the notation $1/A$ for the inverse of the matrix A)

$$\Pi_1 := -\frac{z^{-1}}{1+m_{2c}(z)\Sigma}, \quad \Pi_2 := -\frac{z^{-1}}{1+m_{1c}(z)B}.$$

Note that from (A.12) we can express the Stieltjes transforms m_c and $m_{1,2c}$ (which determine the limiting spectral distribution), as the following weighted traces of the functionals of Π:

$$\begin{aligned} \frac{1}{n}\operatorname{Tr}\Pi_1 = m_c, \quad \frac{1}{n}\operatorname{Tr}(\Sigma\Pi_1) = m_{1c}, \\ \frac{1}{n}\operatorname{Tr}(B\Pi_2) = m_{2c}. \end{aligned} \tag{A.23}$$

In [30], [106], an anisotropic local law away from the support of ρ_c was proved in the form of Theorem A.7 below. Roughly speaking, the local law means that the random resolvent matrix G is well approximated by the deterministic matrix Π defined above. This holds in the sense that linear combinations of entries of G can be approximated by the same linear combinations of entries of Π. This has been more formal in work on deterministic equivalents, see e.g., [37], [60].

Before stating the local law, for convenience, we recall the following notion of stochastic domination, which was introduced in [44] and subsequently used in many works on random matrix theory [12], [13], [14], [45], [46], [70]. It simplifies the presentation of the results and their proofs by systematizing statements of the form "ξ is bounded by ζ with high probability up to a small power of n".

Definition A.5 (Stochastic domination)**.** *(i) Let*

$$\xi = \left(\xi^{(n)}(u) : n\in\mathbb{N}, u\in U^{(n)}\right),$$
$$\zeta = \left(\zeta^{(n)}(u) : n\in\mathbb{N}, u\in U^{(n)}\right),$$

be two families of nonnegative random variables, where $U^{(n)}$ is a possibly n-dependent parameter set. We say ξ is stochastically dominated by ζ, uniformly in u, if for any fixed (small) $\varepsilon>0$ and (large) $D>0$,

$$\sup_{u\in U^{(n)}} \mathbb{P}\left(\xi^{(n)}(u) > n^{\varepsilon}\zeta^{(n)}(u)\right) \leqslant n^{-D}$$

for large enough $n\geqslant n_0(\varepsilon,D)$, and we shall use the notation $\xi\prec\zeta$ or $\xi=\mathrm{O}_\prec(\zeta)$. If for some complex-valued family ξ we have $|\xi|\prec\zeta$, then we will also write $\xi\prec\zeta$.

(ii) Let A be a family of random matrices and ζ be a family of nonnegative random variables. Then $A=\mathrm{O}_\prec(\zeta)$ means that $\|A\|\prec\zeta$.

(iii) We say an event Ξ holds with high probability if for any constant $D>0$, $\mathbb{P}(\Xi^c)\leqslant n^{-D}$ for all large enough n. We say an event Ξ holds with high probability on an event Ω if for any constant $D>0$, $\mathbb{P}(\Omega\backslash\Xi)\leqslant n^{-D}$ for all large enough n.

The following lemma collects basic properties of stochastic domination, which will be used repeatedly in the proof.

Lemma A.6 (Lemma 3.2 in [12], Closure properties of stochastic domination)**.** *Let ξ and ζ be families of nonnegative random variables. Let $C>0$ be any (large) constant.*

- *(i)* **Sums.** *Suppose that $\xi(u,v)\prec\zeta(u,v)$ uniformly in $u\in U$ and $v\in V$. If $|V|\leqslant n^C$, then $\sum_{v\in V}\xi(u,v)\prec\sum_{v\in V}\zeta(u,v)$ uniformly in u.*
- *(ii)* **Products.** *If $\xi_1(u)\prec\zeta_1(u)$ and $\xi_2(u)\prec\zeta_2(u)$ uniformly in $u\in U$, then $\xi_1(u)\xi_2(u)\prec\zeta_1(u)\zeta_2(u)$ uniformly in u.*
- *(iii)* **Taking expectations.** *Suppose that $\Psi(u)\geqslant n^{-C}$ is deterministic and $\xi(u)$ satisfies $\mathbb{E}\xi(u)^2\leqslant n^C$ for all u. Then if $\xi(u)\prec\Psi(u)$ uniformly in u, we have $\mathbb{E}\xi(u)\prec\Psi(u)$ uniformly in u.*

In this paper, given (possibly complex-valued) vectors u,v and a matrix A of conformable dimensions, we denote the inner product by

$$\langle u, Av\rangle := u^\top Av,$$

where $u^\top$ is the complex conjugate of u. For simplicity, we shall also write $\langle u,Av\rangle$ as a generalized entry $A_{uv}:=\langle u,Av\rangle$.

Now we are ready to state the anisotropic local law for G, which will be the main tool of this paper. It essentially follows from Theorem 4.10 of [30]. However, our setting is a little different from the setting there, so we will give the necessary details in Appendix F to adapt the proof in [30] to our setting.

Theorem A.7 (Anisotropic local law outside of the spectrum)**.** *Suppose the setting in Section A-A and Assumption A.3 hold. Let $\mathscr{A}$ be any set of (complex-valued) deterministic unit*

vectors of cardinality $|\mathscr{A}| \leqslant n^C$ for some constant $C > 0$. Fix any small constant $c_0 > 0$ and large constant $C_0 > 0$, and define

$$S_{out}(c_0, C_0) := \{E + \mathrm{i}\eta : \lambda_+ + c_0 \leqslant E \leqslant C_0, \\ \eta \in [0, C_0]\}. \tag{A.24}$$

There exists a set Ω with $\mathbb{P}(\Omega) \geqslant 1 - n^{-\delta}$ for some constant $0 < \delta \leqslant 1/2$ depending on τ in (A.8) *only, such that the following anisotropic local law holds:*

$$1(\Omega) \max_{u,v \in \mathscr{A}} |\langle u, G(X,z)v\rangle - \langle u, \Pi(z)v\rangle| \prec n^{-\delta} \tag{A.25}$$

uniformly in $z \in S_{out}(c_0, C_0)$.

We remark that Theorem 2.4 of [12] is actually a special case of our Theorem A.7 by replacing $\widetilde{X} = S$ (i.e., we replace $X \mapsto S$, $\Sigma \mapsto I_p$ and $S \mapsto I_n$), but on a bigger domain of z. In fact, our Theorem A.7 can be also generalized to such a bigger domain of z by Theorem 3.6 of [106], and the reader can check that (C.6) below holds due to the claim in (A.25).

Moreover, we have the following local law for z near the edge λ_+ of the spectrum, which will be used to study the non-spiked eigenvalues and eigenvectors. It is a corollary of Theorem 3.6 of [106], and we shall give the proof in Appendix F. The only difference between the local law outside the spectrum and the one near the edge is that the argument $z = E + \mathrm{i}\eta$ of the resolvent $G(X, z)$ is restricted to have real part E strictly larger than the right edge λ_+ for the law outside the spectrum, and there are no restrictions on the imaginary part η. For the law near the edge, z is restricted to have real part E around the right edge λ_+, but the imaginary part η must have absolute value at least of the order of $n^{-1/2+c_1}$ for some $c_1 > 0$.

Theorem A.8 (Anisotropic local law near the edge)**.** *Suppose the assumptions of Theorem A.7 hold. Fix any small constants $c_0, c_1 > 0$ and large constant $C_0 > 0$, and define*

$$S_{edge}(c_0, C_0, c_1) := \{E + \mathrm{i}\eta : \lambda_+ - c_0 \leqslant E \leqslant C_0, \\ \eta \in [n^{-1/2+c_1}, C_0]\Big\}. \tag{A.26}$$

There exists a set Ω with $\mathbb{P}(\Omega) \geqslant 1 - n^{-\delta}$ for some constant $0 < \delta \leqslant 1/2$ depending on τ in (A.8) *only, such that the following anisotropic local law holds as long as c_0 is small enough depending on γ, ξ, τ:*

$$1(\Omega) \max_{u,v \in \mathscr{A}} |\langle u, G(X,z)v\rangle - \langle u, \Pi(z)v\rangle| \prec n^{-\delta} \tag{A.27}$$

uniformly in $z \in S_{edge}(c_0, C_0, c_1)$. Moreover, fixing any $\varpi \in \mathbb{N}$, we have that

$$1(\Omega) \max_{1 \leqslant i \leqslant \varpi} |\lambda_i - \lambda_+| \prec n^{-\delta}. \tag{A.28}$$

We mention that such local laws are part of a much broader line of work in random matrix theory, going back to the Marchenko-Pastur law [79]. See e.g., [48], [49] for more recent results on related topics such as universality. The topic of deterministic equivalents is also related, see e.g., [37], [60].

C. The spiked eigenvalues and eigenvectors

With the anisotropic local law, we can derive a so-called master equation for the outlier eigenvalues and eigenvectors. We write the sketched signal matrix as

$$\sum_{i=1}^{k} d_i v_i u_i^\top = V D U^\top, \quad D = \mathrm{diag}(d_1, \cdots, d_k),$$

where U, V and W are $p \times k$, $r \times k$ and $n \times k$ matrices:

$$U = (u_1, \cdots, u_k), \quad V = (v_1, \cdots, v_k) = SW, \\ W = (w_1, \cdots, w_k).$$

Then we define the *linearization of the sketched signal* as the following $(p + r) \times (p + r)$ block matrix:

$$\Delta H := z^{1/2} \begin{pmatrix} 0 & UDV^\top \\ VDU^\top & 0 \end{pmatrix} = z^{1/2} A \mathcal{D} A^\top,$$

where

$$A := \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}, \quad \mathcal{D} := \begin{pmatrix} 0 & D \\ D & 0 \end{pmatrix}.$$

Lemma A.9. *If $x > \lambda_+$ is not an eigenvalue of $\mathcal{Q}_1 = \widetilde{X}^\top \widetilde{X}$, then it is an eigenvalue of $\widetilde{\mathcal{Q}}_1 = \widetilde{Y}^\top \widetilde{Y}$ if and only if the following determinant (of a $2k \times 2k$ matrix) vanishes:*

$$\det\left(\mathcal{D}^{-1} + x^{1/2} A^\top G(x) A\right) = 0. \tag{A.29}$$

Proof. The proof is similar to the one for Lemma 5.1 of [30]. Since our setting is somewhat different, we give a full proof. The non-zero eigenvalues of $z^{-1/2}\widetilde{H}$ are

$$\pm\sqrt{\lambda_1(\widetilde{\mathcal{Q}}_1)}, \ \pm\sqrt{\lambda_2(\widetilde{\mathcal{Q}}_1)}, \ \cdots, \ \pm\sqrt{\lambda_{p\wedge r}(\widetilde{\mathcal{Q}}_1)}.$$

Hence $x > 0$ is an eigenvalue of $\widetilde{\mathcal{Q}}_1$ if and only if

$$\det\left(\widetilde{H}(X, x) - x\right) = 0, \tag{A.30}$$

from which we obtain that

$$\begin{aligned} 0 &= \det(H + \Delta H - x) \\ &= \det(H - x) \det\left(I + G(x)\Delta H\right) \\ &= \det(H - x) \det(I + x^{1/2} A^\top G(x) A \mathcal{D}) \\ &= \det(\mathcal{D}) \det(H - x) \det(\mathcal{D}^{-1} + x^{1/2} A^\top G(x) A), \end{aligned}$$

where in the third step we used identity $\det(I + \mathcal{C}\mathcal{B}) = \det(I + \mathcal{B}\mathcal{C})$ for any two matrices $\mathcal{B}$ and $\mathcal{C}$ of conformable dimensions, and identity matrices of appropriate dimensions. The claim then follows since $\det(H - x) \neq 0$. □

Using Theorem A.7, up to some small error of order $\mathrm{O}_\prec(n^{-\delta})$, equation (A.29) gives approximately the following *eigenvalue master equation* for any possible spike x:

$$\det M(x) := \det \begin{pmatrix} M_1(x) & D^{-1} \\ D^{-1} & M_2(x) \end{pmatrix} = 0, \tag{A.31}$$

where

$$M_1(x) := -x^{-1/2} U^\top \left(1 + m_{2c}(x)\Sigma\right)^{-1} U, \\ M_2(x) := -x^{-1/2} V^\top (1 + m_{1c}(x) B)^{-1} V.$$

In summary, we have the following setup for finding the limiting locatons of the spikes after sketching:

(i) We are given the population covariance Σ ($p \times p$).
(ii) We have the sketching matrix S ($r \times n$).
(iii) We are given the left and right matrices of eigenvectors V and U ($r \times k$ and $p \times k$).
(iv) We have the $k \times k$ diagonal matrix D of population spikes.
(v) For any given x, we calculate the pair $(m_{1c}(x), m_{2c}(x))$, arising as the solution to the self-consistent equations in (A.12).
(vi) We combine the above quantities into the $2k \times 2k$ master matrix $M(x)$ from (A.31).
(vii) We solve for the values x for which this matrix is singular, i.e., solve equation (A.31). In general we expect at most k such values. These are all possible candidates for the limits of the empirical spikes of the sketched data.

To get concrete results, we solve the master equation in special cases as in Section III.

Next we discuss the sample eigenvectors for the outliers (i.e., the spikes). For now, suppose we know that the i-th largest outlier $\widetilde{\lambda}_i$ is close to a "classical location" θ_i, which does not depend on n, p, r, but can depend on the other parameters. Moreover, assume that these values are well-separated from each other (i.e. there exists a constant $\varepsilon > 0$ such that $|\theta_i - \theta_j| \geqslant \varepsilon$ for any $i \neq j$).[1] We want to study the overlap between the sample eigenvector and the population eigenvector u_i. Let

$$\widetilde{Y} = \sum_{k=1}^{p \wedge r} \widetilde{\lambda}_k^{1/2} \widetilde{\zeta}_k \widetilde{\xi}_k^\top,$$

be a singular value decomposition of the sketched spiked matrix, where

$$\widetilde{\lambda}_1 \geqslant \widetilde{\lambda}_2 \geqslant \ldots \geqslant \widetilde{\lambda}_{p \wedge r} \geqslant 0 = \widetilde{\lambda}_{p \wedge r + 1} = \ldots = \widetilde{\lambda}_{p \vee r}$$

are the eigenvalues of $\widetilde{\mathcal{Q}}_1 = \widetilde{Y}^\top \widetilde{Y}$, while $\{\widetilde{\zeta}_k\}_{k=1}^r$ and $\{\widetilde{\xi}_k\}_{k=1}^p$ are the left and right singular vectors of $\widetilde{Y}$, respectively. Then using (A.21) for $\widetilde{G}$, we find that for $i, j \in \mathcal{I}_1$ and $\mu, \nu \in \mathcal{I}_2$,

$$\widetilde{G}_{ij} = \sum_{k=1}^{p} \frac{\widetilde{\xi}_k(i) \widetilde{\xi}_k^\top(j)}{\widetilde{\lambda}_k - z}, \qquad \widetilde{G}_{\mu\nu} = \sum_{k=1}^{r} \frac{\widetilde{\zeta}_k(\mu) \widetilde{\zeta}_k^\top(\nu)}{\widetilde{\lambda}_k - z}, \quad \text{(A.32)}$$

and

$$\begin{aligned} \widetilde{G}_{i\mu} &= z^{-1/2} \sum_{k=1}^{p \wedge r} \frac{\widetilde{\lambda}_k^{1/2} \widetilde{\xi}_k(i) \widetilde{\zeta}_k^\top(\mu)}{\widetilde{\lambda}_k - z}, \\ \widetilde{G}_{\mu i} &= z^{-1/2} \sum_{k=1}^{p \wedge r} \frac{\widetilde{\lambda}_k^{1/2} \widetilde{\zeta}_k(\mu) \widetilde{\xi}_k^\top(i)}{\widetilde{\lambda}_k - z}. \end{aligned} \quad \text{(A.33)}$$

We also recall the following well known lemma, which follows from a simple algebraic calculation.

Lemma A.10 (Woodbury matrix identity)**.** *For $\mathcal{A}, S, \mathcal{B}, T$ of conformable dimensions, we have*

$$\begin{aligned} &(\mathcal{A} + S\mathcal{B}T)^{-1} \\ &= \mathcal{A}^{-1} - \mathcal{A}^{-1} S (\mathcal{B}^{-1} + T\mathcal{A}^{-1} S)^{-1} T \mathcal{A}^{-1}. \end{aligned} \quad \text{(A.34)}$$

[1]In our paper ε denotes a constant, and its value can change at each appearance.

as long as all operations are well defined. As a special case, we have the following equation, sometimes known as Hua's identity:

$$\mathcal{A} - \mathcal{A}(\mathcal{A} + \mathcal{B})^{-1} \mathcal{A} = \mathcal{B} - \mathcal{B}(\mathcal{A} + \mathcal{B})^{-1} \mathcal{B} \quad \text{(A.35)}$$

if $\mathcal{A} + \mathcal{B}$ is non-singular.

With (A.34), we can write that

$$\begin{aligned} A^\top \widetilde{G}(z) A &= A^\top \frac{1}{H - z + z^{1/2} A \mathcal{D} A^\top} A \\ &= A^\top \left(G(z) - G(z) A \frac{1}{z^{-1/2} \mathcal{D}^{-1} + A^\top G(z) A} A^\top G(z) \right) A. \end{aligned}$$

Our goal is to study $|\langle u_j, \widetilde{\xi}_i \rangle|^2$ for some spiked eigenvector $\widetilde{\xi}_i$. We consider a small contour Γ_i around θ_i, which only encloses $\widetilde{\lambda}_i$ and no other eigenvalues. Then using Cauchy's Theorem, we obtain the following *angle master equation*:

$$\begin{aligned} |\langle u_j, \widetilde{\xi}_i \rangle|^2 &= \frac{-1}{2\pi \mathrm{i}} \oint_{\Gamma_i} e_j^\top A^\top \widetilde{G}(z) A e_j \mathrm{d}z \quad \text{(A.36)} \\ &= \frac{1}{2\pi \mathrm{i} (\widetilde{\lambda}_i)^{1/2}} \oint_{\Gamma_i} e_j^\top \mathcal{D}^{-1} \frac{1}{\mathcal{D}^{-1} + z^{1/2} A^\top G(z) A} \mathcal{D}^{-1} e_j \mathrm{d}z. \end{aligned}$$

This gives an expression for the inner product of the true and empirical spike eigenvectors. To evaluate it in specific cases, again we need to study the master matrix $M(z)^{-1} = \left(\mathcal{D}^{-1} + z^{1/2} A^\top G(z) A \right)^{-1}$.

APPENDIX B
PROOF OF THEOREM III.1

In this section, we prove Theorem III.1 based on the master equations (A.31) and (A.36), and the local laws, Theorems A.7 and A.8. We give the details of the proof in this section, which can be applied to Theorems III.2–III.6 directly. The only differences will be the analysis of the master equations, which we will perform in a case by case manner.

In our proof, we will use the following asymptotic notations. Given (n-dependent) quantities A_n and B_n, we write $A_n = \mathrm{O}(B_n)$ or $|A_n| \lesssim |B_n|$ if there exists a constant $C > 0$ such that $|A_n| \leqslant C|B_n|$ for large enough n, and we write $A_n \sim B_n$ if $A_n = \mathrm{O}(B_n)$ and $B_n = \mathrm{O}(A_n)$. We write $A_n = \mathrm{o}(B_n)$ or $|A_n| \ll |B_n|$ to indicate that $|A_n| \leqslant c_n |B_n|$ for a positive sequence of numbers $c_n \downarrow 0$ as $n \to \infty$.

We first introduce some preliminary bounds. For $z = E + \mathrm{i}\eta$, we define the distance to the rightmost edge as

$$\kappa \equiv \kappa_E := |E - \lambda_+|. \quad \text{(B.1)}$$

Then we summarize some basic properties of $m_{1,2c}$. We define the domain

$$\begin{aligned} \widetilde{S}(c_0, C_0) := \{ &z = E + \mathrm{i}\eta : \lambda_+ - c_0 \leqslant E \leqslant C_0, \\ &0 \leqslant \eta \leqslant C_0 \}. \end{aligned} \quad \text{(B.2)}$$

Lemma B.1 (Lemma 3.4 of [106])**.** *Suppose* (A.4)*,* (A.7)*, and Assumption A.3 hold. Fix any constant $C_0 > 0$. Then there exists a sufficiently small constant $c_0 > 0$ depending on γ, ξ, τ such that for $z = E + \mathrm{i}\eta \in \widetilde{S}(c_0, C_0)$,*

(i) for $\alpha = 1, 2$*,*

$$|m_{\alpha c}(z)| \sim 1, \ \operatorname{Im} m_{\alpha c}(z) \sim \begin{cases} \frac{\eta}{\sqrt{\kappa+\eta}}, & \text{if } E \geqslant \lambda_+ \\ \sqrt{\kappa+\eta}, & \text{if } E \leqslant \lambda_+ \end{cases}; \quad \text{(B.3)}$$

(ii) there exists a constant $\tau' > 0$ *such that*

$$\min_{\mu} |1 + m_{1c}(z) s_\mu| \geqslant \tau', \ \min_{i} |1 + m_{2c}(z)\sigma_i| \geqslant \tau'. \quad \text{(B.4)}$$

In fact, (B.4) *holds if we replace* s_μ *(resp.* σ_i*) with any positive value that is smaller than* s_1 *(resp.* σ_1*).*

The functions $m_{1c}(z)$ and $m_{2c}(z)$ are holomorphic on the right half complex plane $\{z : \operatorname{Re} z > \lambda_+\}$. Moreover, they are one-to-one in the region near the real axis, so that we can define their inverse functions g_{1c} and g_{2c}. The following lemma gives some basic bounds on $m_{1,2c}$, $g_{1,2c}$ and their derivatives.

Lemma B.2 (Lemma 4.5 of [30])**.** *Suppose the assumptions of Lemma B.1 hold. Then for any constant* $\varsigma > 0$*, there exist constants* $\tau_0, \tau_1, \tau_2 > 0$ *depending on* $\gamma, \xi, \tau, \varsigma$ *such that the following statements hold.*

(i) m_{1c} *and* m_{2c} *are holomorphic homeomorphisms on the spectral domain*

$$D(\tau_0, \varsigma) := \{z = E + \mathrm{i}\eta : \lambda_+ < E < \varsigma, \ -\tau_0 < \eta < \tau_0\}.$$

As a consequence, the inverse functions of m_{1c} *and* m_{2c} *exist and we denote them by* g_{1c} *and* g_{2c}*, respectively.*

(ii) We have $D_1(\tau_1, \varsigma) \subset m_{1c}(D(\tau_0, \varsigma))$ *and* $D_2(\tau_2, \varsigma) \subset m_{2c}(D(\tau_0, \varsigma))$*, where*

$$D_1(\tau_1, \varsigma) := \{\xi = E + \mathrm{i}\eta : m_{1c}(\lambda_+) < E < m_{1c}(\varsigma), \\ -\tau_1 < \eta < \tau_1\},$$

and

$$D_2(\tau_2, \varsigma) := \{\zeta = E + \mathrm{i}\eta : m_{2c}(\lambda_+) < E < m_{2c}(\varsigma), \\ -\tau_2 < \eta < \tau_2\}.$$

In other words, g_{1c} *and* g_{2c} *are holomorphic homeomorphisms on* $D_1(\tau_1, \varsigma)$ *and* $D_2(\tau_2, \varsigma)$*, respectively.*

(iii) For $z \in D(\tau_0, \varsigma)$*, we have*

$$\begin{aligned} |m_{1c}(z) - m_{1c}(\lambda_+)| &\sim |z - \lambda_+|^{1/2}, \\ |m_{2c}(z) - m_{2c}(\lambda_+)| &\sim |z - \lambda_+|^{1/2}, \end{aligned} \quad \text{(B.5)}$$

and

$$\begin{aligned} |m'_{1c}(z)| &\sim |z - \lambda_+|^{-1/2}, \\ |m'_{2c}(z)| &\sim |z - \lambda_+|^{-1/2}. \end{aligned} \quad \text{(B.6)}$$

(iv) For $z_1, z_2 \in D(\tau_0, \varsigma)$*, we have*

$$\begin{aligned} |m_{1c}(z_1) - m_{1c}(z_2)| &\sim |m_{2c}(z_1) - m_{2c}(z_2)| \\ &\sim \frac{|z_1 - z_2|}{\max_{i=1,2} |z_i - \lambda_+|^{1/2}}. \end{aligned} \quad \text{(B.7)}$$

The following eigenvalue interlacing result follows directly from the Cauchy interlacing theorem.

Lemma B.3 (Eigenvalue interlacing)**.** *Recall that the eigenvalues of* $\widetilde{\mathcal{Q}}_1$ *and* $\mathcal{Q}_1$ *are denoted by* $\{\widetilde{\lambda}_i\}$ *and* $\{\lambda_i\}$*, respectively. Then we have*

$$\widetilde{\lambda}_i \in [\lambda_{i+k}, \lambda_{i-k}], \quad \text{(B.8)}$$

where we adopt the convention that $\lambda_i = \infty$ *if* $i < 1$ *and* $\lambda_i = 0$ *if* $i > p \wedge r$*.*

With the above preparations, we are ready to prove Theorem III.1. We first prove the near-orthogonality of columns of partial orthogonal matrices, (III.9).

Proof of (III.9). Let us represent $\widehat{S}$ as the upper $r \times n$ submatrix of some $n \times n$ Haar distributed matrix T. Then we have

$$\begin{aligned} \mathbb{E}(V^\top V)_{11} &= \mathbb{E} \sum_{j=1}^{r} \widehat{S}_{j1}^2 = \frac{1}{n} \mathbb{E} \sum_{j=1}^{r} \sum_{k=1}^{n} \widehat{S}_{jk}^2 = \frac{r}{n}, \\ \operatorname{Var}\left[(V^\top V)_{11}\right] &= \mathbb{E} \sum_{j,j'=1}^{r} \widehat{S}_{j1}^2 \widehat{S}_{j'1}^2 - \frac{r^2}{n^2} \\ &= \mathbb{E} \sum_{j=1}^{r} \widehat{S}_{j1}^4 + \sum_{j \neq j' \in [\![1,r]\!]} \mathbb{E} \widehat{S}_{j1}^2 \widehat{S}_{j'1}^2 - \frac{r^2}{n^2} \\ &= \mathbb{E} \sum_{j=1}^{r} T_{j1}^4 + \frac{r(r-1)}{n(n-1)} \mathbb{E} \sum_{j \neq j' \in [\![1,n]\!]} T_{j1}^2 T_{j'1}^2 - \frac{r^2}{n^2} \\ &= \mathbb{E} \sum_{j=1}^{r} T_{j1}^4 - \frac{r(r-1)}{n(n-1)} \mathbb{E} \sum_{j=1}^{n} T_{j1}^4 \\ &\quad + \frac{r(r-1)}{n(n-1)} \mathbb{E} \sum_{j,j'=1}^{n} T_{j1}^2 T_{j'1}^2 - \frac{r^2}{n^2} = \mathrm{O}(n^{-1}), \end{aligned}$$

where we used that $\mathbb{E} T_{j1}^4 = \mathrm{O}(n^{-2})$, since the random vector $t_1 := (T_{j1})$ has the same distribution as a normalized Gaussian vector:

$$t_1 \overset{d}{=} g/\|g\|.$$

Here $g \in \mathbb{R}^n$ has iid standard normal entries. Similarly, we can calculate that

$$\begin{aligned} \mathbb{E}(V^\top V)_{12} &= \sum_{j=1}^{r} \mathbb{E} \widehat{S}_{j1} \widehat{S}_{j2} = \frac{r}{n} \mathbb{E} \left(\sum_{k=1}^{n} T_{k1} T_{k2} \right) = 0, \\ \operatorname{Var}\left[(V^\top V)_{12}\right] &= \mathbb{E} \sum_{j,j'=1}^{r} \widehat{S}_{j1} \widehat{S}_{j2} \widehat{S}_{j'1} \widehat{S}_{j'2} \\ &= \frac{1}{n(n-1)} \sum_{j,j'=1}^{r} \mathbb{E} \sum_{k \neq k' \in [\![1,n]\!]} T_{jk} T_{jk'} T_{j'k} T_{j'k'} \\ &= \frac{1}{n(n-1)} \sum_{j,j'=1}^{r} \mathbb{E} \sum_{k,k'=1}^{n} T_{jk} T_{jk'} T_{j'k} T_{j'k'} \\ &\quad - \frac{1}{n(n-1)} \sum_{j,j'=1}^{r} \mathbb{E} \sum_{k=1}^{n} T_{jk}^2 T_{j'k}^2 \leqslant \frac{r}{n(n-1)}. \end{aligned}$$

Then we conclude (III.9) by Chebyshev's inequality. □

In fact, we know that a much stronger bound holds:

$$V^\top V = \xi_n I_k + \mathrm{O}_\prec(n^{-1/2}) \quad \text{(B.9)}$$

using more advanced tools from random matrix theory. Although we will not use such a strong bound in this paper,

it may be helpful to keep in mind that our result can be improved to give much better convergence rates. For example, if the entries of X have finite a-th moment for a constant $a > 4$, then the results in Theorem III.1 can be obtained with an explicit convergence rate $\mathrm{O}_{\prec}(n^{-\delta})$ for $\delta = 1/2 - 2/a$. In particular, if the entries of X have finite moments up to any order (e.g. when the entries of X are sub-Gaussian), then we can get the optimal convergence rate $\mathrm{O}_{\prec}(n^{-1/2})$ using (B.9) in our proof. The behavior of submatrices of random orthogonal matrices has been well studied, see e.g., [64], [65] and references therein. These works study approximation by Gaussian random random variabes, and require more than what we need in this work.

Now we are ready to prove the eigenvalue bounds in Theorem III.1.

Proof of (III.15) *and* (III.17). By Lemma A.9, it is enough to study the behavior of $A^\top G(x)A$. By Theorem A.7 and Theorem A.8, we can choose a high-probability event $\Xi \subset \Omega$, such that the following bounds hold for some constants $c_0, c_1, C_0 > 0$ and any fixed large integer $\varpi \in \mathbb{N}$:

$$\mathbf{1}(\Xi)\|A^\top(G(z) - \Pi(z))A\| \leqslant n^{-\delta/2}, \tag{B.10}$$

for $z \in S_{edge}(c_0, C_0, c_1) \cup S_{out}(c_0, C_0)$;

$$\mathbf{1}(\Xi)\,|\lambda_i - \lambda_+| \leqslant n^{-\delta/2}, \tag{B.11}$$

for $1 \leqslant i \leqslant \varpi$. We remark that the randomness of X only comes into play to ensure that Ξ holds with high probability. The rest of the proof is restricted to Ξ only, and will be entirely deterministic.

We denote $d_c := \sqrt{\gamma_n/\xi_n}$, and define the index sets

$$\mathcal{O}_+ := \{1 \leqslant i \leqslant k : d_i > d_c\}, \tag{B.12}$$

which is the set of the indices of outliers. We also denote $k_+ := |\mathcal{O}_+|$.

Step 1: Our first step is to prove that on Ξ, there are no eigenvalues outside a neighborhood of the classical outlier locations θ_i. For each $1 \leqslant i \leqslant k_+$, we define the permissible interval

$$\mathrm{I}_i \equiv \mathrm{I}_i(\varepsilon) := [\theta_i - \varepsilon, \theta_i + \varepsilon],$$

where ε is a constant that can be arbitrarily small as long as we have

$$\mathrm{I}_i \cap \mathrm{I}_j = \varnothing, \quad i \neq j. \tag{B.13}$$

Moreover, we define the permissible interval $\mathrm{I}_0 \equiv \mathrm{I}_0(\varepsilon) := [0, \lambda_+ + \varepsilon]$ for other eigenvalues, and denote

$$\mathrm{I} := \mathrm{I}_0 \cup \Big(\bigcup_{i \in \mathcal{O}_+} \mathrm{I}_i \Big). \tag{B.14}$$

We claim the following result.

Lemma B.4. *The complement of* I *contains no eigenvalues of* $\widetilde{\mathcal{Q}}_1$.

Proof. By (A.29), (B.10) and (B.11), we see that $x \notin \mathrm{I}_0$ is an eigenvalue of $\widetilde{\mathcal{Q}}_1$ if and only if

$$\begin{aligned} &\mathcal{D}^{-1} + x^{1/2}A^\top G(x)A \\ &= \mathcal{D}^{-1} + x^{1/2}A^\top \Pi(x)A + \mathrm{O}(n^{-\delta/2}) \end{aligned} \tag{B.15}$$

is singular. By (B.11), we know on Ξ, $\widetilde{\lambda}_1 \leqslant (\sqrt{\lambda_1} + d_1)^2 \leqslant C_0$ as long as C_0 is taken large enough. Here we used the trivial bound for the operator norms,

$$\widetilde{\lambda}_1^{1/2} = \|Y\| \leqslant \|X\| + d_1 = \sqrt{\lambda_1} + d_1.$$

Moreover, by (III.9), we have that with probability $1 - \mathrm{o}(1)$,

$$\begin{aligned} &\mathcal{D}^{-1} + x^{1/2}A^\top \Pi(x)A \\ &= \begin{pmatrix} -x^{-1/2}\,(1 + m_{2c}(x))^{-1} I_k & D^{-1} \\ D^{-1} & x^{1/2} m_{2c}(x) I_k \end{pmatrix} + \mathrm{o}(1) \end{aligned}$$

for all $x \in [0, C_0]\backslash \mathrm{I}$. Thus to prove the lemma, it suffices to show that if $x \in [0, C_0]\backslash \mathrm{I}$, then

$$\left| \frac{m_{2c}(x)}{1 + m_{2c}(x)} + d_i^{-2} \right| \geqslant c, \quad 1 \leqslant i \leqslant k, \tag{B.16}$$

for some constant $c > 0$ depending only on ε. If (B.16) holds, then we immediately obtain that

$$\left\| \begin{pmatrix} -x^{-1/2}\,(1 + m_{2c}(x))^{-1} I_k & D^{-1} \\ D^{-1} & x^{1/2} m_{2c}(x) I_k \end{pmatrix}^{-1} \right\| = \mathrm{O}(1),$$

and hence $(\mathcal{D}^{-1} + x^{1/2}A^\top G(x)A)$ must be non-singular. This means that x cannot be an eigenvalue of $\widetilde{Q}_1$.

For the proof of (B.16), by (III.11), θ_i satisfy

$$\frac{m_{2c}(\theta_i)}{1 + m_{2c}(\theta_i)} = -d_i^{-2}, \quad 1 \leqslant i \leqslant k_+.$$

Thus we have that for $1 \leqslant i \leqslant k_+$,

$$\begin{aligned} \left| \frac{m_{2c}(x)}{1 + m_{2c}(x)} + d_i^{-2} \right| &= \left| \frac{m_{2c}(x)}{1 + m_{2c}(x)} - \frac{m_{2c}(\theta_i)}{1 + m_{2c}(\theta_i)} \right| \\ &\gtrsim |m_{2c}(x) - m_{2c}(\theta_i)| \gtrsim 1, \end{aligned}$$

where we used (B.4) in the second step and (B.7) in the last step. Moreover, using $0 > m_{2c}(x) \geqslant m_{2c}(\lambda_+ + \varepsilon) > -1$ for $x \in [0, C_0]\backslash \mathrm{I}$ and $m_{2c}(\lambda_+) = -(1 + d_c^2)^{-1}$, we find that for $k_+ \leqslant i \leqslant k$,

$$\begin{aligned} \frac{m_{2c}(x)}{1 + m_{2c}(x)} + d_i^{-2} &\geqslant \frac{m_{2c}(x)}{1 + m_{2c}(x)} + d_c^{-2} \gtrsim m_{2c}(x) + \frac{1}{1 + d_c^2} \\ &\geqslant m_{2c}(\lambda_+ + \varepsilon) - m_{2c}(\lambda_+) \gtrsim 1, \end{aligned}$$

where again we used (B.4) in the second step and (B.7) in the last step. This concludes (B.16), which further proves Lemma B.4. □

Step 2: In this step, we claim the following result.

Lemma B.5. *Each* I_i, $1 \leqslant i \leqslant k_+$, *contains precisely one eigenvalue of* $\widetilde{\mathcal{Q}}_1$.

Proof. Fix any $1 \leqslant i \leqslant k_+$ and pick a sufficiently small positively oriented closed contour $\mathcal{C} \subset \mathbb{C}/[0, \lambda_+]$ that encloses θ_i but no other point of the set $\{\theta_i\}_{i=1}^{k_+}$. By (B.13), we can choose the contour $\mathcal{C}$ as a circle around θ_i with radius ε.

Now we define two functions

$$\begin{aligned} h(z) &:= \det(\mathcal{D}^{-1} + z^{1/2}A^\top G(z)A), \\ l(z) &:= \det(\mathcal{D}^{-1} + z^{1/2}A^\top \Pi(z)A). \end{aligned}$$

By (B.11), the functions h, l are holomorphic on and inside $\mathcal{C}$ when n is sufficiently large. Moreover, by the construction

of $\mathcal{C}$, the function l has precisely one zero inside $\mathcal{C}$; at θ_i. By (B.10) and a similar argument as for (B.16), we have

$$\min_{z\in\mathcal{C}}|l(z)| \gtrsim 1, \quad |h(z)-l(z)| = \mathrm{O}(n^{-\delta/2}).$$

The lemma then follows from Rouché's theorem. □

Combining Steps 1 and 2 with a simple eigenvalue counting argument, we obtain that

$$1(\Xi)|\widetilde{\lambda}_i - \theta_i| \leqslant \varepsilon, \quad 1 \leqslant i \leqslant k_+, \tag{B.17}$$

and

$$1(\Xi)\widetilde{\lambda}_i \leqslant 1(\Xi)\lambda_+ + \varepsilon, \quad k_+ \leqslant i \leqslant k, \tag{B.18}$$

for any small constant $\varepsilon > 0$. The first bound (B.17) concludes (III.15). To prove (III.17), we still need to provide a lower bound for $\widetilde{\lambda}_i$, $k_+ \leqslant i \leqslant k$. In fact, with (B.8) and (B.11), we obtain that

$$1(\Xi)\widetilde{\lambda}_i \geqslant 1(\Xi)\lambda_+ - n^{-\delta/2}, \quad k_+ \leqslant i \leqslant k.$$

Together with (B.18), we conclude (III.17). □

Finally we prove the eigenvector bounds in Theorem III.1.

Proof of (III.16) *and* (III.18). In the following proof, we again always work on the event Ξ such that (B.10) and (B.11) hold. Again the randomness of X only comes into play to ensure that Ξ holds with high probability, and the rest of the proof is deterministic on Ξ.

For $\mathcal{E}(z) = z^{1/2}A^\top(\Pi(z) - G(z))A$, we have $z^{1/2}A^\top G(z)A = z^{1/2}A^\top\Pi(z)A - \mathcal{E}(z)$. By (B.10), we have that

$$\|\mathcal{E}(z)\| \leqslant n^{-\delta/2} \quad \text{for} \quad z \in S_{out}(c_0, C_0). \tag{B.19}$$

We now perform a resolvent expansion for the denominator in (A.36) as

$$\begin{aligned}\frac{1}{\mathcal{D}^{-1} + z^{1/2}A^\top G(z)A} &= \frac{1}{\mathcal{D}^{-1} + z^{1/2}A^\top\Pi(z)A} \\ &+ \frac{1}{\mathcal{D}^{-1} + z^{1/2}A^\top\Pi(z)A}\mathcal{E}\frac{1}{\mathcal{D}^{-1} + z^{1/2}A^\top G(z)A}.\end{aligned} \tag{B.20}$$

We define the contour $\Gamma_i = \{z : |z-\theta_i| = \varepsilon\}$, where $\varepsilon > 0$ is a sufficiently small constant such that

$$\inf_{z\in\Gamma_i}\left(|z-\lambda_+| \wedge \min_{1\leqslant i\leqslant k_+}|z-\theta_i|\right) \geqslant \varepsilon. \tag{B.21}$$

By (III.15) and (III.17), for large enough n, we have (i) Γ_i only encloses $\widetilde{\lambda}_i$, and no other eigenvalue of $\widetilde{\mathcal{Q}}_1$; (ii) Γ_i does not enclose any pole of G (i.e. any eigenvalue of $\mathcal{Q}_1$). Note (i) implies that Γ_i only encloses one pole of $(\mathcal{D}^{-1} + z^{1/2}A^\top G(z)A)^{-1}$ at $\widetilde{\lambda}_i$. Moreover, with a similar argument as for (B.16), one can obtain that

$$\max_{z\in\Gamma_i}\|(\mathcal{D}^{-1} + z^{1/2}A^\top\Pi(z)A)^{-1}\| \leqslant c^{-1}$$

for some constant $c > 0$ depending on ε only. Together with (B.19), we find that

$$\begin{aligned}&\max_{z\in\Gamma_i}\left\|\frac{1}{\mathcal{D}^{-1} + z^{1/2}A^\top\Pi(z)A}\mathcal{E}\frac{1}{\mathcal{D}^{-1} + z^{1/2}A^\top G(z)A}\right\| \\ &\lesssim n^{-\delta/2}.\end{aligned} \tag{B.22}$$

Now inserting (B.20) into (A.36), choosing Γ_i as above, and using (B.22), we obtain from Cauchy's integral formula that for $1 \leqslant i \leqslant k_+$ and $1 \leqslant j \leqslant k$,

$$\begin{aligned}|\langle u_j, \widetilde{\xi}_i\rangle|^2 &= \mathrm{o}(1) + \frac{\delta_{ij}}{2\pi\mathrm{i}(\widetilde{\lambda}_i)^{1/2}}\oint_{\Gamma_i}(0, d_i^{-1}) \\ &\times \begin{pmatrix} -z^{-1/2}(1+m_{2c}(z))^{-1} & d_i^{-1} \\ d_i^{-1} & z^{1/2}m_{2c}(z)\end{pmatrix}^{-1}\begin{pmatrix}0\\ d_i^{-1}\end{pmatrix}\mathrm{d}z \\ &= \frac{\delta_{ij}}{2\pi\mathrm{i}\theta_i(1+d_i^2)}\oint_{\Gamma_i}\frac{1}{m_{2c}(z) + (1+d_i^2)^{-1}}\mathrm{d}z + \mathrm{o}(1) \\ &= \frac{1}{\theta_i(1+d_i^2)m'_{2c}(\theta_i)} + \mathrm{o}(1) \\ &= \delta_{ij}\frac{g'_{2c}(-(1+d_i^2)^{-1})}{\theta_i(1+d_i^2)} + \mathrm{o}(1) \\ &= \frac{\xi_n - \frac{\gamma_n}{d_i^4}}{\xi_n + \frac{\gamma_n}{d_i^2}} + \mathrm{o}(1).\end{aligned} \tag{B.23}$$

where we used (III.15) in the second and third steps, and (III.8) in the last step. This concludes (III.16).

Next we prove (III.18). For $k_+ \leqslant i \leqslant k$, we choose a specific spectral parameter as $z_i = \widetilde{\lambda}_i + \mathrm{i}\eta_i$, where $\eta_i := n^{-\varepsilon}$ for some sufficiently small constant $\varepsilon > 0$. Note that by (III.17), we have $z_i \in S_{edge}(c_0, C_0, c_1)$. With the spectral decomposition (A.32), we obtain that

$$\operatorname{Im}\widetilde{G}_{uu}(E+\mathrm{i}\eta) = \sum_{j=1}^{p}\frac{\eta|\langle u, \widetilde{\xi}_j\rangle|^2}{(\widetilde{\lambda}_j - E)^2 + \eta^2},$$

which implies

$$|\langle u, \widetilde{\boldsymbol{\xi}}_i\rangle|^2 \leqslant \eta_i \operatorname{Im}\langle u, \widetilde{G}(z_i)u\rangle. \tag{B.24}$$

Applying (A.34) to $\widetilde{G}(z_i) = (H + z_i^{1/2}A\mathcal{D}A^\top - z_i)^{-1}$, we obtain that

$$\begin{aligned}\langle u, \widetilde{G}(z_i)u\rangle &= G_{uu}(z_i) \\ &- z_i^{1/2}u^\top G(z_i)A\frac{1}{\mathcal{D}^{-1} + z_i^{1/2}A^\top G(z_i)A}A^\top G(z_i)u.\end{aligned} \tag{B.25}$$

For the denominator, we claim that for sufficiently small ε,

$$\left\|\left(\mathcal{D}^{-1} + z^{1/2}A^\top G(z_i)A\right)^{-1}\right\| \lesssim (\operatorname{Im} m_{2c}(z_i))^{-1}. \tag{B.26}$$

To prove this claim, we first notice that

$$\begin{aligned}\left|\frac{m_{2c}(z_i)}{1+m_{2c}(z_i)} + d_i^{-2}\right| &\gtrsim \left|m_{2c}(z_i) + \frac{1}{1+d_i^2}\right| \\ &\gtrsim \operatorname{Im} m_{2c}(z_i) \gtrsim \eta_i,\end{aligned} \tag{B.27}$$

where we used (B.4) in the second step and (B.3) in the last step. This shows that the smallest singular value of

$$M(z_i) = \mathcal{D}^{-1} + z^{1/2}A^\top\Pi(z_i)A$$

is at least of order $\gtrsim \operatorname{Im} m_{2c}(z_i)$. Then by (B.10), we have that

$$\mathcal{D}^{-1} + z^{1/2}A^\top G(z_i)A = M(z_i) + \mathrm{O}(n^{-\delta/2}).$$

Thus as long as we choose $\varepsilon < \delta/2$, the bound (B.26) holds.

Now using (B.10), we get

$$G_{uu}(z_i) = \mathrm{O}(1), \quad \|u^\top G(z_i) A\| = \mathrm{O}(1).$$

Together with (B.25) and (B.26), we obtain that

$$\eta_i \operatorname{Im}\langle u, \widetilde{G}(z_i)u\rangle \lesssim \frac{\eta_i}{\operatorname{Im} m_{2c}(z_i)} \lesssim \max\{\sqrt{\eta_i}, \sqrt{\kappa_{\widetilde{\lambda}_i}}\}.$$

where $\kappa_{\widetilde{\lambda}_i} = |\widetilde{\lambda}_i - \lambda_+|$ (recall (B.1)) and in the last step we used

$$\operatorname{Im} m_{2c}(\eta_i) \gtrsim \min\left\{\sqrt{\eta_i}, \frac{\eta_i}{\sqrt{\kappa_{\widetilde{\lambda}_i}}}\right\}$$

by (B.3). Hence with $\eta_i = n^{-\varepsilon}$ and $\kappa_{\widetilde{\lambda}_i} = \mathrm{o}(1)$ by (III.17), we conclude from (B.24) that

$$\left|\langle u, \widetilde{\boldsymbol{\xi}}_i\rangle\right|^2 \leqslant \eta_i \operatorname{Im}\langle u, \widetilde{G}(z_i)u\rangle = \mathrm{o}(1).$$

This completes the proof of (III.18). □

APPENDIX C
PROOF OF THEOREM III.2

In this section, we prove Theorem III.2. For the reader's convenience, we first provide an informal argument, and then the fully rigorous proof.

To simplify the notation, we will sometimes denote $A+z := A + zI_q$ for a $q \times q$ matrix A and a complex scalar z. We will need the following resolvents of S (compare with Definition A.1):

$$\begin{aligned} \mathcal{R}_1(S,z) &:= \left(SS^\top - z\right)^{-1}, \\ \mathcal{R}_2(S,z) &:= \left(S^\top S - z\right)^{-1}, \end{aligned} \tag{C.1}$$

and the normalized traces

$$m_1^S(z) := \frac{1}{p}\operatorname{Tr}\mathcal{R}_1(z), \quad m_2^S(z) := \frac{1}{n}\operatorname{Tr}\mathcal{R}_2(z). \tag{C.2}$$

Let m_{1c}^S and m_{2c}^S be the limiting Stieltjes transforms of $SS^\top$ and $S^\top S$. They are determined by the following self-consistent equations:

$$\begin{aligned} m_{1c}^S(z) &= \frac{1}{-z\left[1 + m_{2c}^S(z)\right]}, \\ m_{2c}^S(z) &= \frac{1}{-z\left[1 + \xi_n m_{1c}^S(z)\right]}. \end{aligned} \tag{C.3}$$

Solving (C.3), we can obtain that

$$\begin{aligned} m_{1c}^S(z) &= \frac{-(z-1+\xi_n) + \sqrt{(z-\lambda_+^S)(z-\lambda_-^S)}}{2z\xi_n}, \\ m_{2c}^S(z) &= \frac{-(z+1-\xi_n) + \sqrt{(z-\lambda_+^S)(z-\lambda_-^S)}}{2z}, \end{aligned} \tag{C.4}$$

where $\lambda_\pm^S$ are the edges of the support of the standard Marchenko-Pastur (MP) distribution,

$$\lambda_\pm^S = (1 \pm \sqrt{\xi_n})^2.$$

Denoting the inverse functions of $m_{1,2c}^S$ by $g_{1,2c}^S$, we also obtain from the equations in (C.3) that

$$g_{1c}^S(m) = \frac{1}{1+\xi_n m} - \frac{1}{m}, \quad g_{2c}^S(m) = \frac{\xi_n}{1+m} - \frac{1}{m}. \tag{C.5}$$

By the *local law* for isotropic sample covariance matrices, Theorem 2.4 of [12], we know that for any deterministic unit vectors $u_1, u_2 \in \mathbb{R}^r$ and $v_1, v_2 \in \mathbb{R}^n$,

$$\begin{aligned} \langle u_1, \mathcal{R}_1(z) u_2\rangle &= m_{1c}^S(z)\langle u_1, u_2\rangle + \mathrm{o}(1), \\ \langle v_1, \mathcal{R}_2(z) v_2\rangle &= m_{2c}^S(z)\langle v_1, v_2\rangle + \mathrm{o}(1), \end{aligned} \tag{C.6}$$

with high probability, uniformly in the following region of z bounded away from the support of the MP law:

$$S_\tau := \{z \in \mathbb{C} : \operatorname{dist}(z, [(1-\sqrt{\xi_n})^2, (1+\sqrt{\xi_n})^2]) \geqslant \tau\} \tag{C.7}$$

for any constant $\tau > 0$. In particular, (C.6) implies that

$$m_1^S(z) = m_{1c}^S(z) + \mathrm{o}(1), \quad m_2^S(z) = m_{2c}^S(z) + \mathrm{o}(1), \tag{C.8}$$

uniformly in $z \in S_\tau$.

Now using m_1^S, we can write the self-consistent equations in (A.12) as

$$\begin{aligned} m_{1c}(z) &= \gamma_n \frac{1}{-z\left[1 + m_{2c}(z)\right]}, \\ m_{2c}(z) &= \frac{\xi_n}{-z m_{1c}(z)}\left(1 - \frac{1}{m_{1c}(z)} m_1^S(-m_{1c}^{-1}(z))\right). \end{aligned} \tag{C.9}$$

Suppose that (C.8) can be applied to m_1^S. Then we obtain the following self-consistent equation satisfied by m_{1c}:

$$\begin{aligned} &\frac{\gamma_n}{m_{1c}(z)} = -z \\ &+ \frac{\xi_n}{m_{1c}(z)}\left(1 - \frac{1}{m_{1c}(z)} m_{1c}^S(-m_{1c}^{-1}(z))\right) + \mathrm{o}(1). \end{aligned} \tag{C.10}$$

This immediately gives the inverse function g_{1c} of m_{1c}:

$$g_{1c}(m) = -\frac{\gamma_n}{m} + \frac{\xi_n}{m}\left(1 - \frac{1}{m} m_{1c}^S(-m^{-1})\right) + \mathrm{o}(1). \tag{C.11}$$

Next we find the function $m_{1c}(z)$. Using the function g_{1c}^S as an inverse function of m_{1c}^S, we can obtain that m_{1c} in (C.10) satisfies (approximately) the following equation:

$$\begin{aligned} -\frac{1}{m_{1c}} &= g_{1c}^S\left(-\frac{\gamma_n - \xi_n}{\xi_n} m_{1c} - \frac{z}{\xi_n} m_{1c}^2\right) \\ &= \frac{1}{1 - (\gamma_n - \xi_n) m_{1c} - z m_{1c}^2} \\ &\quad + \frac{\xi_n}{(\gamma_n - \xi_n) m_{1c} + z m_{1c}^2}, \end{aligned}$$

which can be reduced to a cubic equation

$$\begin{aligned} &z^2 m_{1c}^3 - z(1 + \xi_n - 2\gamma_n) m_{1c}^2 \\ &- \left(z + (1-\gamma_n)(\gamma_n - \xi_n)\right) m_{1c} - \gamma_n = 0. \end{aligned} \tag{C.12}$$

There is only one solution to this equation such that $\operatorname{Im} m_{1c}(z) > 0$ whenever $\operatorname{Im} z > 0$. After obtaining $m_{1c}(z)$, we immediately obtain that the Stieltjes transform m, the limit of $\frac{1}{p}\operatorname{Tr}(Y^\top Y - z)^{-1}$, has the form

$$m(z) = m_c(z) + \mathrm{o}(1), \quad m_c(z) := \xi_n^{-1} m_{1c}(z),$$

with high probability. Hence we can define the asymptotic spectral density ρ_c using the inversion formula $\rho_c(E) = \lim_{\eta \downarrow 0} \pi^{-1} \operatorname{Im} m_c(E + \mathrm{i}\eta)$, and find its right edge λ_+.

To study the spiked eigenvalues and eigenvectors, we again need to study the master matrix $M(x)$ in (II.3). With the Woodbury matrix identity (A.34), we obtain that

$$\begin{aligned}&-x^{-1/2}W^\top S^\top\frac{1}{1+m_{1c}SS^\top}SW\\&=-\frac{1}{x^{1/2}m_{1c}}W^\top\left(1-\frac{1}{1+m_{1c}S^\top S}\right)W.\end{aligned}$$

Applying the local law (C.6), we obtain that with high probability,

$$\begin{aligned}&-x^{-1/2}W^\top S^\top\frac{1}{1+m_{1c}SS^\top}SW\\&=-\frac{1}{x^{1/2}m_{1c}}\left(1-\frac{1}{m_{1c}}m_{2c}^S(-m_{1c}^{-1})\right)I_k+\mathrm{o}(1).\end{aligned}$$

Now the eigenvalue master equation (A.31) becomes, approximately,

$$\det\begin{pmatrix}\gamma_n^{-1}m_{1c}I_k & D^{-1}\\ D^{-1} & \frac{-1}{m_{1c}}\left(1-\frac{1}{m_{1c}}m_{2c}^S(-m_{1c}^{-1})\right)I_k\end{pmatrix}=0,$$

which gives the following equations for $1\leqslant i\leqslant k$,

$$-\gamma_n^{-1}\left(1-\frac{1}{m_{1c}}m_{2c}^S(-m_{1c}^{-1})\right)=d_i^{-2}.$$

Using the inverse function of m_{2c}^S, g_{2c}^S, in (C.5), we obtain that

$$-m_{1c}^{-1}=g_{2c}^S\left((1+\gamma_nd_i^{-2})m_{1c}\right),$$

which gives

$$m_{1c}(x)=-\frac{\gamma_nd_i^{-2}}{\left(1+\gamma_nd_i^{-2}\right)\left(\xi_n+\gamma_nd_i^{-2}\right)}. \tag{C.13}$$

Similarly to (III.12), in order for the signal strength d_i to give an outlier, we need to have that

$$\alpha(d_i)=-\frac{\gamma_nd_i^{-2}}{\left(1+\gamma_nd_i^{-2}\right)\left(\xi_n+\gamma_nd_i^{-2}\right)}>m_{1c}(\lambda_+). \tag{C.14}$$

In particular, there exists a $d_c>0$ determined by the equation

$$\alpha(d_c)=m_{1c}(\lambda_+), \tag{C.15}$$

such that (C.14) holds if and only if $d_i>d_c$. If $d_i>d_c$, then the i-th outlier $\widetilde{\lambda}_i$ will concentrate around

$$\theta_i=g_{1c}\left(-\frac{\gamma_nd_i^{-2}}{\left(1+\gamma_nd_i^{-2}\right)\left(\xi_n+\gamma_nd_i^{-2}\right)}\right) \tag{C.16}$$

by (C.13), where g_{1c} is defined in (C.11).

Next we study the spiked eigenvector corresponding to the outlier $\widetilde{\lambda}_i$ using the angle master equation (A.36). First it is easy to see that $|\langle u_j,\widetilde{\xi}_i\rangle|^2=\mathrm{o}(1)$ if $j\neq i$. If $j=i$, then we have that with high probability,

$$\begin{aligned}&|\langle u_i,\widetilde{\xi}_i\rangle|^2\\&=\frac{1}{2\pi\mathrm{i}\sqrt{\theta_i}d_i^2}\oint_{\Gamma_i}\frac{-z^{1/2}\gamma_n^{-1}m_{1c}(z)}{\gamma_n^{-1}\left(1-\frac{1}{m_{1c}(z)}m_{2c}^S(-m_{1c}^{-1}(z))\right)+d_i^{-2}}\mathrm{d}z\\&\quad+\mathrm{o}(1)\\&=-\frac{m_{1c}^2(\theta_i)}{2\pi\mathrm{i}d_i^2}\oint_{\Gamma_i}\frac{1}{(1+\gamma_nd_i^{-2})m_{1c}(z)-m_{2c}^S(-m_{1c}^{-1}(z))}\mathrm{d}z\\&\quad+\mathrm{o}(1)\\&=-\frac{m_{1c}^2(\theta_i)}{2\pi\mathrm{i}d_i^2}\oint_{g_{1c}(\Gamma_i)}\frac{g_{1c}'(\zeta)}{(1+\gamma_nd_i^{-2})\zeta-m_{2c}^S(-\zeta^{-1})}\mathrm{d}\zeta+\mathrm{o}(1)\\&=\frac{\alpha_i^2}{d_i^2}\frac{g_{1c}'(\alpha_i)}{\alpha_i^{-2}(m_{2c}^S)'(-\alpha_i^{-1})-(1+\gamma_nd_i^{-2})}+\mathrm{o}(1),\end{aligned} \tag{C.17}$$

where we used that $\alpha_i\equiv\alpha(d_i)=m_{1c}(\theta_i)+\mathrm{o}(1)$.

One can see that in order to make the above calculations rigorous, we only need to repeat the arguments in the proof for Theorem III.1, with two additional steps. (i) We need to verify that the "right edge regularity" condition (A.16) holds, such that the anisotropic local law *outside* the spectrum (Theorem A.7) can be applied. (ii) We need to verify that $-m_{1c}^{-1}(z)\in S_\tau$ for some constant $\tau>0$ for all $z\in S_{out}(c_0,C_0)\cup S_{edge}(c_0,C_0,c_1)$, such that we can apply (C.6).

Proof of Theorem III.2. We first verify the condition (A.16). The self-consistent equation (A.12) now becomes

$$\begin{aligned}-zm_{1c}(z)&=\frac{\gamma_n}{1+m_{2c}(z)},\\-zm_{2c}(z)&=\frac{1}{n}\sum_{\mu=1}^r\frac{s_\mu}{1+s_\mu m_{1c}(z)}.\end{aligned} \tag{C.18}$$

By the last statement of Lemma A.2, the two sums on the right-hand side of the above two equations are all positive sums if we take $z=\lambda_+$. If $1+m_{2c}(\lambda_+)=\mathrm{o}(1)$, then from the first equation we find that $|m_{1c}(\lambda_+)|\gg1$, which contradicts the fact that $|m_{1c}(\lambda_+)|\leqslant s_1^{-1}$.

On the other hand, suppose

$$1+m_{1c}(\lambda_+)s_1=\mathrm{o}(1). \tag{C.19}$$

From (C.18), we obtain the following self-consistent equation for m_{2c}:

$$f(m_{2c}(z))=0,$$

where

$$f(m_{2c}):=m_{2c}-\frac{1}{n}\sum_{\mu=1}^r\frac{s_\mu(1+m_{2c})}{-z(1+m_{2c})+s_\mu\gamma_n}.$$

If we regard f as a function of m_{2c}, then by Lemma 2.5 of [106], we know that $\partial_{m_{2c}}f=0$ at $z=\lambda_+$. Hence we get

$$\begin{aligned}1&=\frac{1}{n}\sum_{\mu=1}^r\frac{s_\mu^2\gamma_n}{[-\lambda_+(1+m_{2c}(\lambda_+))+s_\mu\gamma_n]^2}\\&=\frac{m_{1c}^2(\lambda_+)}{n}\sum_{\mu=1}^r\frac{s_\mu^2\gamma_n^{-1}}{[1+s_\mu m_{1c}(\lambda_+)]^2}.\end{aligned} \tag{C.20}$$

By the eigenvalue rigidity result for $SS^\top$, Theorem 2.10 of [12] or Theorem 3.8 of [106], we know that for any small constant $\varepsilon>0$,

$$\max_{1\leqslant\mu\leqslant\varepsilon n}|s_\mu-s_1|\leqslant C\varepsilon^{2/3}\quad\text{with high probability,}$$

for some constant $C > 0$ that is independent of ε. Combining it with the hypothesis (C.19), we obtain from (C.20) that with high probability,

$$m_{1c}^2(\lambda_+) \leqslant C\varepsilon^{1/3}$$

for some constant $C > 0$ that is independent of ε. However, this contradicts (C.19) if we take ε to be sufficiently small. Thus (A.16) holds with high probability.

Next we show that $-m_{1c}^{-1}(z) \in S_\tau$ for some constant $\tau > 0$ for all $z \in \widetilde{S}(c_0, C_0)$ (recall (B.2)) as long as c_0 is sufficiently small. Again by the eigenvalue rigidity result for $SS^\top$, Theorem 2.10 of [12], we know that $|s_1 - (1+\sqrt{\xi_n})^2| = \mathrm{o}(1)$ with high probability. Hence together with (A.16), we have

$$-m_{1c}^{-1}(\lambda_+) \geqslant (1+\sqrt{\xi_n})^2 + c_1$$

for some constant $c_1 > 0$ depending on τ. Moreover, since $m_{1c}(\lambda_+) \leqslant m_{1c}(x) < 0$ for $x > \lambda_+$ and $m_{1c}(x)$ is monotonically increasing in $x \in (\lambda_+, \infty)$, we obtain that

$$\inf_{x \geqslant \lambda_+} \left[-m_{1c}^{-1}(x)\right] \geqslant (1+\sqrt{\xi_n})^2 + c_1.$$

Next if $\mathrm{dist}(z, [\lambda_+, C_0]) \leqslant \delta$ for some constant $\delta > 0$, by (B.5) we obtain that

$$\inf_{z:\mathrm{dist}(z,[\lambda_+,C_0])\leqslant\delta} \min\{1 + s_\mu m_{1c}(z)\} \geqslant c_1/2$$

as long as δ is taken sufficiently small. If we take $c_0 \leqslant \delta$, the above bound covers all the domain $\widetilde{S}(c_0, C_0)$ except for the part $\{z \in \widetilde{S}(c_0, C_0) : \mathrm{Im}\, z \geqslant c_0\}$. On this part of domain, we use (B.3) to get that

$$\begin{aligned}&\inf_{z\in\widetilde{S}(c_0,C_0):\mathrm{Im}\, z\geqslant c_0} \left[-m_{1c}^{-1}(z)\right] \\ &\geqslant c_2' \inf_{z\in\widetilde{S}(c_0,C_0):\mathrm{Im}\, z\geqslant c_0} \mathrm{Im}\, m_{1c}(z) \geqslant c_2\end{aligned}$$

for some constants $c_2, c_2' > 0$ depending on c_0.

In sum, we find that $-m_{1c}^{-1}(z) \in S_{\tau'}$ for some constant $\tau' \equiv \tau'_{c_1,c_2} > 0$ for all $z \in \widetilde{S}(c_0, C_0) \supset S_{out}(c_0, C_0) \cup S_{edge}(c_0, C_0, c_1)$.

The above proof justifies our calculations between (C.9) and (C.17). The rest of the proof is exactly the same as the one for Theorem III.1, so we omit the details. □

APPENDIX D
PROOF OF THEOREM III.3 AND THEOREM III.5

As remarked at the beginning of Appendix B, we only need to analyze the master equations (A.31) and (A.36) under the settings of Theorem III.3 and Theorem III.5, respectively. The rest of the proof will be exactly the same as the one for Theorem III.1 in Appendix B.

Proof of Theorem III.3. We define the random variable

$$\widetilde{\xi}_n := \frac{1}{n}\sum_{i=1}^n S_{ii}$$

to be the fraction of non-zero diagonal entries of S. We fix a realization of S. Then the equations in (A.12) become

$$\begin{aligned}m_{1c}(z) &= \gamma_n \frac{1}{-z\left[1 + m_{2c}(z)\right]}, \\ m_{2c}(z) &= \widetilde{\xi}_n \frac{1}{-z\left[1 + m_{1c}(z)\right]}.\end{aligned} \tag{D.1}$$

Thus (III.7) and (III.8) are accurate asymptotically, since $\widetilde{\xi}_n$ concentrates around ξ_n for large r and n. Indeed, we can calculate that

$$\begin{aligned}&- x^{-1/2} W^\top S^\top (1 + m_{1c}(x) SS^\top)^{-1} SW \\ &= -\frac{1}{x^{1/2}(1 + m_{1c}(x))} W^\top S^2 W.\end{aligned}$$

This equality holds because $SS^\top = S^2$ is a diagonal matrix with 0-1 entries. Now under the assumption (III.24), we claim that

$$W^\top S^2 W \to \xi_n I_k \quad \text{in probability.} \tag{D.2}$$

Again this follows from a simple moment calculation. We can calculate that

$$\mathbb{E}\sum_{l=1}^n S_{ll}^2 w_i(l) w_j(l) = \frac{r}{n}\sum_{l=1}^n w_i(l) w_j(l) = \frac{r}{n}\delta_{ij}.$$

Then we can calculate the variances: for $i \neq j$,

$$\begin{aligned}&\mathbb{E}\left(\sum_{l=1}^n S_{ll}^2 w_i(l) w_j(l)\right)^2 \\ &= \mathbb{E}\sum_{l=1}^n \varepsilon_l w_i^2(l) w_j^2(l) + \mathbb{E}\sum_{l\neq l'} \varepsilon_l \varepsilon_{l'} w_i(l) w_i(l') w_j(l) w_j(l') \\ &\leqslant \frac{r}{n}\sum_{l=1}^n w_i^2(l) w_j^2(l) + \left(\frac{r}{n}\right)^2 \sum_{l,l'=1}^n w_i(l) w_i(l') w_j(l) w_j(l') \\ &= \frac{r}{n}\sum_{l=1}^n w_i^2(l) w_j^2(l) \leqslant \|w_i\|_\infty^2 \to 0,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\left(\sum_{l=1}^n \left(S_{ll}^2 - \frac{r}{n}\right) w_i^2(l)\right)^2 \\ &= \frac{r}{n}\left(1 - \frac{r}{n}\right)\sum_{l=1}^n w_i^4(l) \leqslant \|w_i\|_\infty^2 \to 0.\end{aligned}$$

Hence (D.2) holds and we obtain that

$$\begin{aligned}&- x^{-1/2} W^\top S^\top (1 + m_{1c}(x) SS^\top)^{-1} SW \\ &\to_P -x^{-1/2}\frac{\xi_n}{1 + m_{1c}} I_k = x^{1/2} m_{2c}(x) I_k.\end{aligned}$$

Hence the matrix $M(x)$ takes the same form as the one in the uniform orthogonal random projection case in Section III-A, which concludes Theorem III.3 with the same arguments as in Appendix B. □

Proof of Theorem III.5. In general, we can write

$$B = \widehat{S}\widehat{S}^\top = \mathrm{diag}(\widehat{c}_1, \cdots, \widehat{c}_r), \quad \widehat{c}_i := 1_{c_i > 0}.$$

Let $\widehat{r}$ be the random number of nonzero c_i-s, and denote

$$\widehat{\xi}_n := \xi_n \left[1 - \exp(-1/\xi_n)\right].$$

By the Poisson convergence theorem, we have

$$\frac{\widehat{r}}{n} = \xi_n \left[1 - P(Poisson(1/\xi_n) = 0)\right] + \mathrm{o}(1) = \widehat{\xi}_n + \mathrm{o}(1)$$

in probability. Thus the self-consistent equation (A.12) becomes

$$\begin{aligned} m_{1c}(z) &= \gamma_n \frac{1}{-z\left[1 + m_{2c}(z)\right]}, \\ m_{2c}(z) &= \widehat{\xi}_n \frac{1}{-z\left[1 + m_{1c}(z)\right]} + \mathrm{o}(1) \end{aligned} \tag{D.3}$$

in probability. We claim that under the delocalization condition (III.24),

$$W^\top \widehat{S}^\top \widehat{S} W = \widehat{\xi}_n I_k + \mathrm{o}(1) \quad \text{in probability.} \tag{D.4}$$

If (D.4) holds, then we have

$$\begin{aligned} &W^\top \widehat{S}^\top (1 + m_{1c}(x)\widehat{S}\widehat{S}^\top)^{-1} \widehat{S} W \\ &= \frac{1}{1 + m_{1c}(x)} W^\top \widehat{S}^\top \widehat{S} W = \frac{\widehat{\xi}_n}{1 + m_{1c}(x)} + \mathrm{o}(1) \end{aligned}$$

in probability, which shows that the master matrix has the following form

$$M(x) = \begin{pmatrix} \frac{-x^{-1/2}}{1+m_{2c}(x)} I_k & D^{-1} \\ D^{-1} & x^{1/2} m_{2c}(x) I_k \end{pmatrix} + \mathrm{o}(1)$$

in probability. Again $M(x)$ takes the same form as in the uniform orthogonal random projection case in Section III-A, except that we replace ξ_n with $\widehat{\xi}_n$. Then one can conclude Theorem III.5 with the same arguments in Appendix B.

It remains to prove the concentration claim (D.4). We again calculate the means and variances. Note that for any vector v,

$$(\widehat{S}v)(i) = 1_{c_i > 0} \frac{1}{\sqrt{c_i}} \sum_{i:h(j)=i} S_{h(j),j} v(j).$$

For $1 \leqslant \alpha, \beta \leqslant k$, we have

$$\begin{aligned} &\left(W^\top \widehat{S}^\top \widehat{S} W\right)_{\alpha\beta} \\ &= \sum_{i=1}^r \frac{1_{c_i>0}}{c_i} \sum_{j:h(j)=i} a_j w_\alpha(j) \cdot \sum_{j':h(j')=i} a_{j'} w_\beta(j'). \end{aligned}$$

We first calculate the mean. Notice that the following conditional expectation can be calculated exactly as

$$\mathbb{E}\left[\sum_{j:h(j)=i} w_\alpha(j) w_\beta(j) \middle| c_i\right] = \frac{c_i}{n} \delta_{\alpha\beta}, \tag{D.5}$$

because by definition $\{j : h(j) = i\}$ is a randomly chosen subset of size c_i, and the vectors w_α-s are orthonormal. Applying (D.5), we get

$$\begin{aligned} &\mathbb{E}\left(W^\top \widehat{S}^\top \widehat{S} W\right)_{\alpha\beta} \\ &= \mathbb{E}\left\{\sum_{i=1}^r \frac{1_{c_i>0}}{c_i} \mathbb{E}\left[\sum_{j:h(j)=i} w_\alpha(j) w_\beta(j) \middle| c_i\right]\right\} \\ &= \mathbb{E}\left\{\sum_{i=1}^r \frac{1_{c_i>0}}{c_i} \frac{c_i}{n} \delta_{\alpha\beta}\right\} = \left(\widehat{\xi}_n + \mathrm{o}(1)\right) \delta_{\alpha\beta}. \end{aligned}$$

Then for $\alpha \neq \beta$, we have

$$\begin{aligned} &\mathbb{E}\left|\left(W^\top \widehat{S}^\top \widehat{S} W\right)_{\alpha\beta}\right|^2 \\ &= \mathbb{E}\left\{\sum_{i_1,i_2=1}^r \frac{1_{c_{i_1}>0, c_{i_2}>0}}{c_{i_1} c_{i_2}} \sum_{h(j_1)=h(j_1')=i_1} \sum_{h(j_2)=h(j_2')=i_2} \right. \\ &\qquad \left. a_{j_1} a_{j_1'} a_{j_2} a_{j_2'} w_\alpha(j_1) w_\beta(j_1') w_\alpha(j_2) w_\beta(j_2')\right\} \\ &= \mathbb{E}\left\{\sum_{i_1 \neq i_2} \frac{1_{c_{i_1}>0, c_{i_2}>0}}{c_{i_1} c_{i_2}} \sum_{j_1:h(j_1)=i_1} w_\alpha(j_1) w_\beta(j_1) \right. \\ &\qquad \left. \times \sum_{j_2:h(j_2)=i_2} w_\alpha(j_2) w_\beta(j_2)\right\} \\ &+ \mathbb{E} \sum_{i=1}^r \frac{1_{c_i>0}}{c_i^2} \sum_{j_1 \neq j_2 : h(j_1)=h(j_2)=i} 2 w_\alpha(j_1) w_\beta(j_1) w_\alpha(j_2) w_\beta(j_2) \\ &+ \mathbb{E}\left\{\sum_{i=1}^r \frac{1_{c_i>0}}{c_i^2} \sum_{j_1,j_2:h(j_1)=h(j_2)=i} w_\alpha(j_1)^2 w_\beta(j_2)^2\right\} \\ &= -\mathbb{E}\left\{\sum_{i_1 \neq i_2} \frac{1_{c_{i_1}>0, c_{i_2}>0}}{c_{i_2}(n - c_{i_2})} \left(\sum_{j_2:h(j_2)=i_2} w_\alpha(j_2) w_\beta(j_2)\right)^2\right\} \\ &\quad + \mathrm{o}(1) = \mathrm{o}(1). \end{aligned}$$

Here in the third step we used (III.24) to get that

$$\begin{aligned} &\sum_{i=1}^r \frac{1_{c_i>0}}{c_i^2} \sum_{j_1 \neq j_2 : h(j_1)=h(j_2)=i} 2 |w_\alpha(j_1) w_\beta(j_1) w_\alpha(j_2) w_\beta(j_2)| \\ &+ \sum_{i=1}^r \frac{1_{c_i>0}}{c_i^2} \sum_{j_1,j_2:h(j_1)=h(j_2)=i} w_\alpha(j_1)^2 w_\beta(j_2)^2 \\ &\leqslant 2 \|w_\alpha\|_\infty \|w_\beta\|_\infty \sum_{i=1}^r \sum_{j_1:h(j_1)=i} |w_\alpha(j_1) w_\beta(j_1)| \\ &\quad + \|w_\beta\|_\infty^2 \sum_{i=1}^r \sum_{j_1:h(j_1)=i} |w_\alpha(j_1)|^2 \\ &\leqslant 2 \|w_\alpha\|_\infty \|w_\beta\|_\infty + \|w_\beta\|_\infty^2 = \mathrm{o}(1), \end{aligned}$$

and a similar result as in (D.5), that is, given c_{i_1} and the j_2-s such that $h(j_2) = i_2$,

$$\begin{aligned} &\mathbb{E}\Big[\sum_{j_1:h(j_1)=i_1} w_\alpha(j_1) w_\beta(j_1) \Big| h^{-1}(i_2)\Big] \\ &= \frac{c_{i_1}}{n - c_{i_2}} \sum_{j:h(j)\neq i_2} w_\alpha(j) w_\beta(j) \\ &= -\frac{c_{i_1}}{n - c_{i_2}} \sum_{j:h(j)=i_2} w_\alpha(j) w_\beta(j). \end{aligned}$$

With similar methods, we can calculate the variance of $(W^\top \widehat{S}^\top \widehat{S} W)_{\alpha\alpha}$:

$$\begin{aligned} &\mathbb{E}\left|\left(W^\top \widehat{S}^\top \widehat{S} W\right)_{\alpha\alpha}\right|^2 \\ &= \mathbb{E}\left\{\sum_{i_1,i_2=1}^r \frac{1_{c_{i_1}>0, c_{i_2}>0}}{c_{i_1} c_{i_2}} \sum_{h(j_1)=h(j_1')=i_1} \sum_{h(j_2)=h(j_2')=i_2}\right. \end{aligned}$$

$$
\begin{aligned}
&a_{j_1}a_{j_1'}a_{j_2}a_{j_2'}w_\alpha(j_1)w_\alpha(j_1')w_\alpha(j_2)w_\alpha(j_2')\Big\}\\
&= \mathbb{E}\left\{\sum_{i_1\neq i_2}\frac{1_{c_{i_1}>0,c_{i_2}>0}}{c_{i_1}c_{i_2}}\sum_{j_1:h(j_1)=i_1}w_\alpha(j_1)^2\right.\\
&\qquad\left.\times\sum_{j_2:h(j_2)=i_2}w_\alpha(j_2)^2\right\}\\
&+\mathbb{E}\left\{\sum_{i=1}^r\frac{1_{c_i>0}}{c_i^2}\sum_{j_1\neq j_2:h(j_1)=h(j_2)=i_1}3w_\alpha(j_1)^2w_\alpha(j_2)^2\right\}\\
&+\mathbb{E}\left\{\sum_{i=1}^r\frac{1_{c_i>0}}{c_i^2}\sum_{j_1:h(j_1)=i}w_\alpha(j_1)^4\right\}\\
&= \mathbb{E}\left\{\sum_{i_1\neq i_2}\frac{1_{c_{i_1}>0,c_{i_2}>0}}{c_{i_2}(n-c_{i_2})}\left(1-\mathrm{O}(c_{i_2}\|w_\alpha\|_\infty^2)\right)\right.\\
&\qquad\left.\times\sum_{j_2:h(j_2)=i_2}w_\alpha(j_2)^2\right\}+\mathrm{o}(1)\\
&= \widehat{\xi}_n\mathbb{E}\left\{\sum_{i_2}\frac{1_{c_{i_2}>0}}{c_{i_2}}\sum_{j_2:h(j_2)=i_2}w_\alpha(j_2)^2\right\}+\mathrm{o}(1)\\
&= \widehat{\xi}_n^2+\mathrm{o}(1)=\left|\mathbb{E}\left(W^\top\widehat{S}^\top\widehat{S}W\right)_{\alpha\alpha}\right|^2+\mathrm{o}(1),
\end{aligned}
$$

where in the third step we again used (III.24), and that given c_{i_1} and the j_2-s such that $h(j_2)=i_2$,

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{j_1:h(j_1)=i_1}w_\alpha(j_1)^2\Big|h^{-1}(i_2)\Big]\\
&=\frac{c_{i_1}}{n-c_{i_2}}\sum_{j:h(j)\neq i_2}w_\alpha(j)^2\\
&=\frac{c_{i_1}}{n-c_{i_2}}\Big[1-\sum_{j:h(j)=i_2}w_\alpha(j)^2\Big].
\end{aligned}
$$

Combining the above results with Chebyshev's inequality, we conclude the concentration result (D.4), which further concludes Theorem III.5. □

Appendix E
Proof of Theorem III.6

In the following proof, we only consider the uniform random projection. However, as we have already seen in Section III, the same result also holds for uniform random sampling under the delocalization condition (III.24), for randomized Hadamard sampling, and for CountSketch under the delocalization condition (III.24) but with ξ replaced by $\widehat{\xi}$.

Now we study the i-th spiked eigenvalue and its eigenvector under the assumption (III.30) for some large enough constant $C_0>0$. First, the self-consistent equations in (A.12) become

$$
\begin{aligned}
m_{1c}(z)&=\gamma_n\int\frac{x}{-z\left[1+xm_{2c}(z)\right]}\pi_\Sigma(dx),\\
m_{2c}(z)&=\frac{\xi_n}{-z\left[1+m_{1c}(z)\right]},
\end{aligned}
\tag{E.1}
$$

which are generalizations of (III.5) with $\Sigma=I$. If θ_i is the classical location for the largest eigenvalue, then we have $\theta_i\sim d_i^2$ and

$$
0<-m_{1,2c}(\theta_i)=-\int\frac{\mathrm{d}\rho_{1,2c}(x)}{x-\theta_i}\sim d_i^{-2}.
$$

Then for x around θ_i, we study the master matrix in (A.31) (up to an $\mathrm{o}(1)$ error in probability)

$$
\begin{aligned}
M(x)&=\begin{pmatrix}-x^{-1/2}U^\top\left(1+m_{2c}(x)\Sigma\right)^{-1}U & D^{-1}\\ D^{-1} & x^{1/2}m_{2c}(x)\end{pmatrix}\\
&=\begin{pmatrix}-x^{-1/2}\left(1-m_{2c}(x)E+\mathrm{O}(l_i^{-2})\right) & D^{-1}\\ D^{-1} & x^{1/2}m_{2c}(x)\end{pmatrix},
\end{aligned}
$$

where recall that $E=U^\top\Sigma U$. Then using Schur complement formula, we obtain

$$
\begin{aligned}
&\det M(x)\\
&=\det\left(-1+m_{2c}(x)E+\mathrm{O}(l_i^{-2})\right)\\
&\quad\times\det\left(m_{2c}(x)+D^{-1}\left(1+m_{2c}(x)E+\mathrm{O}(l_i^{-2})\right)D^{-1}\right)\\
&=\det\left(-1+m_{2c}(x)E+\mathrm{O}(l_i^{-2})\right)\det\left(m_{2c}(x)D^{-2}\right)\\
&\quad\times\det\left((D^2+E)+\mathrm{O}(l_i^{-1})+m_{2c}^{-1}(x)\right).
\end{aligned}
$$

By standard results from perturbation theory (e.g., [96]), we know that the first order perturbation of the i-th eigenvalue of D^2+E is given by $d_i^2+E_{ii}+\mathrm{O}(l_i^{-1})$. Hence, by solving $\det\left((D^2+E)+m_{2c}^{-1}(x)\right)$ for $\theta_i=x$, we get

$$
\begin{aligned}
\theta_i&=g_{2c}\left(-\frac{1}{d_i^2+E_{ii}}+\mathrm{O}(l_i^{-3})\right)\\
&=g_{2c}\left(-\frac{1}{d_i^2+E_{ii}}\right)+\mathrm{O}(l_i^{-1})
\end{aligned}
\tag{E.2}
$$

in probability, where we recall that g_{2c} is the inverse function of m_{2c}.

Then we consider the corresponding eigenvectors: using the Schur complement formula, we find (up to an $\mathrm{o}(1)$ error in probability)

$$
\begin{aligned}
|\langle u_j,\widetilde{\xi}_i\rangle|^2&=\frac{1}{2\pi\mathrm{i}\sqrt{\theta_i}}\oint_{\Gamma_i}z^{-1/2}e_j^\top D^{-1}\\
&\times\frac{1}{m_{2c}(z)+D^{-1}\left(1+m_{2c}(z)E+\mathrm{O}(l_i^{-2})\right)D^{-1}}D^{-1}e_j\mathrm{d}z\\
&=\frac{1}{2\pi\mathrm{i}\theta_i}\oint_{\Gamma_i}e_j^\top\frac{1}{m_{2c}(z)(D^2+E+\mathrm{O}(l_i^{-2}))+1}e_j\mathrm{d}z.
\end{aligned}
$$

Again, standard perturbation theory (e.g., [96]) tells us that the eigenvector of D^2+E up to the first order perturbation is given by

$$
e_i+\sum_{j\neq i,1\leqslant j\leqslant k}e_j\frac{E_{ji}}{d_i^2-d_j^2}.
$$

Thus we find that in probability,

$$
|\langle u_i,\widetilde{\xi}_i\rangle|^2=\frac{g_{2c}'\left(-\alpha_i\right)\alpha_i}{g_{2c}(-\alpha_i)}+\mathrm{O}(l_i^{-2}),
\tag{E.3}
$$

and for $j\neq i$,

$$
|\langle u_j,\widetilde{\xi}_i\rangle|^2=\frac{g_{2c}'\left(-\alpha_i\right)\alpha_i}{g_{2c}(-\alpha_i)}\left|\frac{E_{ji}}{d_i^2-d_j^2}\right|^2+\mathrm{O}(l_i^{-3}),
\tag{E.4}
$$

where

$$\alpha_i := (d_i^2 + E_{ii} + \mathrm{O}(l_i^{-1}))^{-1}.$$

It remains to study the expression for g_{2c}. From (E.1), we obtain that

$$z = \gamma_n \int \frac{x}{1 + x m_{2c}(z)} \pi_\Sigma(\mathrm{d}x) - \frac{\xi_n}{m_{2c}(z)},$$

which gives

$$g_{2c}(m) = \gamma_n \int \frac{x}{1 + xm} \pi_\Sigma(\mathrm{d}x) - \frac{\xi_n}{m}.$$

Then we can calculate that

$$\begin{aligned} \theta_i &= g_{2c}\left(-\frac{1}{d_i^2 + E_{ii}}\right) + \mathrm{O}(l_i^{-1}) \\ &= \xi_n (d_i^2 + E_{ii}) + \gamma_n \rho_1 + \mathrm{O}(l_i^{-1}), \end{aligned} \tag{E.5}$$

and

$$\begin{aligned} \frac{g_{2c}'(-\alpha_i)\,\alpha_i}{g_{2c}(-\alpha_i)} &= \frac{\xi_n - \gamma_n \int \frac{x^2 \alpha_i^2}{(1 - x\alpha_i)^2} \pi_\Sigma(dx)}{\xi_n + \gamma_n \int \frac{x\alpha_i}{1 - x\alpha_i} \pi_\Sigma(dx)} \\ &= \frac{\xi_n - \gamma_n \left(\alpha_i^2 \rho_2\right)}{\xi_n + \gamma_n (\alpha_i \rho_1 + \alpha_i^2 \rho_2)} + \mathrm{O}(l_i^{-3}) \\ &= \frac{\xi_n - \frac{\gamma_n}{d_i^4} \rho_2}{\xi_n + \frac{\gamma_n}{d_i^2}\left(\rho_1 + d_i^{-2}(\rho_2 - \rho_1 E_{ii})\right)} + \mathrm{O}(l_i^{-3}), \end{aligned} \tag{E.6}$$

where ρ_i are the moments of the ESD of Σ (recall (III.32)).

With the above calculations, the rest of the proof is exactly the same as the one for Theorem III.1 in Appendix B. We omit the details.

APPENDIX F
PROOF OF THEOREM A.7 AND THEOREM A.8

In this section, we provide the necessary details to complete the proof of Theorem A.7 and Theorem A.8 based on the results in [106] and [30].

We use a standard cutoff argument. We choose the constant $\delta > 0$ small enough such that $\left(n^{1/2-\delta}\right)^{4+\tau} \geqslant n^{2+\delta}$. Then we introduce the following truncation

$$\widehat{X} := 1_\Omega X, \quad \Omega := \left\{ \max_{i,j} |x_{ij}| \leqslant n^{-\delta} \right\}. \tag{F.1}$$

By the moment conditions (A.8) and a simple union bound, we have

$$\mathbb{P}(\widehat{X} \neq X) = \mathrm{O}(n^{-\delta}). \tag{F.2}$$

Using (A.8) and integration by parts, it is easy to verify that

$$\begin{aligned} \mathbb{E}\,|x_{ij}|\,1_{|x_{ij}| > n^{-\delta}} &= \mathrm{O}(n^{-2-\delta}), \\ \mathbb{E}\,|x_{ij}|^2\,1_{|x_{ij}| > n^{-\delta}} &= \mathrm{O}(n^{-2-\delta}), \end{aligned}$$

which imply that

$$|\mathbb{E}\widehat{x}_{ij}| = \mathrm{O}(n^{-2-\delta}), \quad \mathbb{E}|\widehat{x}_{ij}|^2 = n^{-1} + \mathrm{O}(n^{-2-\delta}). \tag{F.3}$$

Moreover, we clearly have

$$\mathbb{E}|\widehat{x}_{ij}|^4 \leqslant \mathbb{E}|x_{ij}|^4 = \mathrm{O}(n^{-2}).$$

We define the following centered version of $\widehat{X}$: $Z = \widehat{X} - \mathbb{E}\widehat{X}$. Then we have the following proposition for the resolvent $G(W, z)$.

Proposition F.1. *Suppose the assumptions of Theorem A.8 hold and define Z as above. Then we have*

$$\max_{u,v \in \mathscr{A}} |\langle u, G(Z,z) v\rangle - \langle u, \Pi(z) v\rangle| \prec n^{-\delta} \tag{F.4}$$

uniformly in $z \in S_{edge}(c_0, C_0, c_1)$. Moreover, we have that for any fixed $\varpi \in \mathbb{N}$,

$$\max_{1 \leqslant i \leqslant \varpi} |\lambda_i - \lambda_+| \prec n^{-\delta}, \tag{F.5}$$

where λ_i denotes the i-th largest eigenvalue of $\mathcal{Q}_1(Z) := (SZ\Sigma^{1/2})^\top SZ\Sigma^{1/2}$.

Proof. The bounds (F.4) and (F.5) has essentially been proved in Theorem 3.6 and Theorem 3.8 of [106], respectively. The only difference is that in [106], the entries of the random matrix Z all have variances n^{-1}, while in the current case we have

$$\mathbb{E}|Z_{ij}|^2 = n^{-1} + \mathrm{O}(n^{-2-\delta}).$$

However, one can check that the error $\mathrm{O}(n^{-2-\delta})$ is sufficiently small such that it is negligible at each step of the proof in [106], which concludes (F.4) and (F.5). We remark that the first author proved stronger results in Theorem 3.6 and Theorem 3.8 of [106], but they are not necessary for our purpose in this paper. □

Then we show that $G(\widehat{X}, z)$ is sufficiently close to $G(Z, z)$ in the sense of the anisotropic local law.

Proposition F.2. *The bound* (F.4) *holds uniformly for $G(\widehat{X}, z)$ in $z \in S_{edge}(c_0, C_0, c_1)$. Moreover,* (F.5) *holds with high probability for $\lambda_i(\mathcal{Q}_1(\widehat{X}))$.*

Proof. We write $\widehat{X} = Z + \mathbb{E}\widehat{X}$, where by (F.3), we have

$$\max_{i,\mu} |\mathbb{E}\widehat{X}_{i\mu}| = \mathrm{O}(n^{-2-\delta}). \tag{F.6}$$

In particular, this gives that $\|\mathbb{E}\widehat{X}\| \leqslant \|\mathbb{E}\widehat{X}\|_F = \mathrm{O}(n^{-1-\delta})$, which implies (F.5) for $\lambda_i(\mathcal{Q}_1(\widehat{X}))$.

For (F.4), we abbreviate $G(\widehat{X}, z) \equiv \widehat{G}$ and $G(Z, z) \equiv G$. Then it suffices to show that for any deterministic unit vectors u, v,

$$\left| \langle u, \widehat{G}(z) v\rangle - \langle u, G(z) v\rangle \right| \prec n^{-\delta} \tag{F.7}$$

uniformly in $z \in S_{edge}(c_0, C_0, c_1)$. We can write that

$$\widehat{G}(z) = \left(G^{-1}(z) + V\right)^{-1},$$

where

$$V := z^{1/2} \begin{pmatrix} 0 & \Sigma^{1/2} (\mathbb{E}\widehat{X})^\top S^\top \\ S(\mathbb{E}\widehat{X})\Sigma^{1/2} & 0 \end{pmatrix}.$$

Then we expand G using the resolvent expansion

$$\widehat{G} = G - GV\widehat{G}. \tag{F.8}$$

Using the spectral decomposition for $\widehat{G}$ as in (A.32) and (A.33), one can easily see that the following deterministic bound holds:

$$\|\widehat{G}(z)\| = \mathrm{O}(\eta^{-1}).$$

Then we can bound the second part on the right-hand side of (F.8) as

$$|\langle u, GV\widehat{G}v\rangle| \lesssim \eta^{-1}\left(\sum_{a\in\mathcal{I}}|\langle u, GVe_a\rangle|^2\right)^{-1/2} \prec \eta^{-1}\left(\sum_{a\in\mathcal{I}}\sum_{b\in\mathcal{I}}|V_{ba}|^2\right)^{-1/2} \lesssim n^{-1/2-\delta}.$$

where e_a denotes the standard unit vector along the a-th direction, and in the second step we applied (F.4) to $\langle u, GVe_a\rangle = G_{uw}$ by taking $w := Ve_a$, and in the third step we used that $\|V\|_F \lesssim \|\mathbb{E}\widehat{X}\|_F \lesssim n^{-1-\delta}$. This concludes (F.4) for $G(\widehat{X}, z)$. □

Finally we show that (A.25) holds for $G(\widehat{X}, z)$ for z with imaginary part down to the real axis in the spectral domain $S_{out}(c_0, C_0)$.

Proposition F.3. *The bound* (F.4) *holds uniformly for* $G(\widehat{X}, z)$ *in* $z \in S_{out}(c_0, C_0)$.

Proof. In this proof, we abbreviate $G(\widehat{X}, z) \equiv G$ and use the notation $\widetilde{X} = S\widehat{X}\Sigma^{1/2}$ as in (A.3). It remains to show that (F.4) holds for $z = E + \mathrm{i}\eta \in S_{out}(c_0, C_0)$ with $\eta \leqslant n^{-1/2+c_1}$. We denote $\eta_0 := n^{-1/2+c_1}$ and $z_0 := E + \mathrm{i}\eta_0$. With (F.4) at z_0, it suffices to prove that

$$\langle u, (\Pi(z) - \Pi(z_0))\, v\rangle \prec n^{-1/2+c_1}, \tag{F.9}$$

and

$$\langle u, (G(z) - G(z_0))\, v\rangle \prec n^{-\delta}. \tag{F.10}$$

With (B.4), to prove (F.9) it is enough to show that

$$|m_{1c}(z) - m_{1c}(z_0)| + |m_{2c}(z) - m_{2c}(z_0)| \prec n^{-1/2+c_1}, \tag{F.11}$$

which follows immediately from (B.7).

For (F.10), we write $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ and $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, and in the following proof, we will always identify vectors v_1 and v_2 with their embeddings $\begin{pmatrix} v_1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ v_2 \end{pmatrix}$, respectively. Let

$$\widetilde{X} = \sum_{k=1}^{p\wedge r}\sqrt{\lambda_k}\zeta_k\xi_k^\top$$

be the singular value decomposition of $\widetilde{X}$. We shall use (A.32)-(A.33) with $\widetilde{G}$ replaced by G. First, the upper left block gives that

$$|\langle u_1, (G(z) - G(z_0))\, v_1\rangle| \leqslant \sum_{k=1}^{p}\frac{\eta_0|\langle v_1, \xi_k\rangle|^2}{\left[(E-\lambda_k)^2+\eta^2\right]^{1/2}\left[(E-\lambda_k)^2+\eta_0^2\right]^{1/2}}. \tag{F.12}$$

By (F.5), we have for any k, $E - \lambda_k \geqslant E - \lambda_1 \geqslant c_0/2 \gg \eta_0$ with high probability for $z \in S_{out}(c_0, C_0)$. Hence we can bound (F.12) by

$$\begin{aligned}
&|\langle u_1, (G(z) - G(z_0))\, v_1\rangle| \\
&\lesssim \sum_{k=1}^{p}\frac{\eta_0|\langle u_1, \xi_k\rangle|^2}{(E-\lambda_k)^2+\eta_0^2} + \sum_{k=1}^{p}\frac{\eta_0|\langle v_1, \xi_k\rangle|^2}{(E-\lambda_k)^2+\eta_0^2} \\
&= \mathrm{Im}\langle u_1, \sum_{k=1}^{p}\frac{\xi_k\xi_k^\top}{E-z_0}u_1\rangle + \mathrm{Im}\langle v_1, \sum_{k=1}^{p}\frac{\xi_k\xi_k^\top}{E-z_0}v_1\rangle \\
&= \mathrm{Im}\, G_{u_1u_1}(z_0) + \mathrm{Im}\, G_{v_1v_1}(z_0) \\
&\prec n^{-\delta} + \mathrm{Im}\,\Pi_{u_1u_1}(z_0) + \mathrm{Im}\,\Pi_{v_1v_1}(z_0) \prec n^{-\delta},
\end{aligned}$$

where in the fourth step we used (F.4) for $G(z_0)$, and in the last step we used (A.22), (B.4) and (B.3) to get

$$\mathrm{Im}\,\Pi_{u_1u_1}(z_0) + \mathrm{Im}\,\Pi_{v_1v_1}(z_0) \prec \eta_0.$$

Similarly, for the upper right block we have

$$\begin{aligned}
&|\langle u_1, (G(z) - G(z_0))\, v_2\rangle| \\
&\prec \left|z^{-1/2} - (z_0z^{-1})^{1/2}\right||\langle u_1, G(z_0)v_2\rangle| \\
&\quad + \sum_{k=1}^{p\wedge n}\frac{\eta_0\,|\langle u_1, \xi_k\rangle\langle\zeta_k, v_2\rangle|}{|\lambda_k - z||\lambda_k - z_0|} \\
&\prec \eta_0 + \sum_{k=1}^{p\wedge n}\frac{\eta_0\,|\langle u_1, \xi_k\rangle|^2}{|\lambda_k - z_0|^2} + \sum_{k=1}^{p\wedge n}\frac{\eta_0\,|v_2, \zeta_k\rangle|^2}{|\lambda_k - z_0|^2} \\
&= \eta_0 + \mathrm{Im}\, G_{u_1u_1}(z_0) + \mathrm{Im}\, G_{v_2v_2}(z_0) \prec n^{-\delta}.
\end{aligned}$$

The lower left and lower right blocks can be handled in the same way. This proves (F.10), which completes the proof. □

Finally, with Proposition F.2, Proposition F.3, and the definition of the truncation (F.1), we conclude Theorem A.7 and Theorem A.8.

APPENDIX G
EXTENSION TO CENTERED MODEL

In this section, we explain how to extend our results to centered sample covariance matrices. We will only consider the following model

$$\widetilde{Y}_a = \widetilde{X}_a + (I - ee^\top)\sum_{i=1}^{k} d_iv_iu_i^\top,$$

where

$$\widetilde{X}_a = (I - ee^\top)SX\Sigma^{1/2}.$$

The other model

$$\widetilde{Y}_b := S(I - ee^\top)X\Sigma^{1/2} + \sum_{i=1}^{k} d_iS(I - ee^\top)w_iu_i^\top$$

can be studied with exactly the same method.

Our goal is to study the principal components of $\widetilde{\mathcal{Q}}_1^a := \widetilde{Y}_a^\top\widetilde{Y}_a$ using the methods in Section A-C. Then we have the following claim.

Claim G.1. *If*

$$\max_i|e^\top Sw_i| = \mathrm{o}(1), \tag{G.1}$$

and uniformly in $z \in S_{out}(c_0, C_0)$,

$$\max_i\left|e^\top\frac{1}{1 + m_{1c}(z)SS^\top}Sw_i\right| = \mathrm{o}(1), \tag{G.2}$$

then the spike eigenvalues and eigenvectors of $\widetilde{\mathcal{Q}}_1^a$ *have the same asymptotic behavior as those of* $\widetilde{\mathcal{Q}}_1 = \widetilde{Y}^\top\widetilde{Y}$.

Proof. Note that under (G.1), we have

$$\|ee^\top \sum_{i=1}^k d_i v_i u_i^\top\| = \|ee^\top S \sum_{i=1}^k d_i w_i u_i^\top\| = \mathrm{o}(1).$$

Then by (A.31) and (A.36), it suffices to show that the same local law holds for $G_a(z)$: with high probability in Ω,

$$\max_{1\leqslant i,j\leqslant k} |\langle Sw_i, [G_a(z) - \Pi(z)]\, Sw_j\rangle| = \mathrm{o}(1) \tag{G.3}$$

uniformly in $z \in S_{out}(c_0, C_0)$, where

$$G_a(z) := \left[z^{1/2} \begin{pmatrix} 0 & \widetilde{X}_a^\top \\ \widetilde{X}_a & 0 \end{pmatrix} - z \right]^{-1}.$$

We denote $S_a := (I - ee^\top)S$, and $B_a := S_a S_a^\top$ with eigenvalues $s_1^a \geqslant s_2^a \geqslant \ldots \geqslant s_r^a \geqslant 0$. Then we can define m_{1c}^a and m_{2c}^a using the self-consistent equation (A.12) by replacing π_B with π_{B_a}, and define $\Pi_a(z)$ by replacing B with B_a, and $m_{1,2c}$ with $m_{1,2c}^a$. We claim that

$$|m_{1c}^a - m_{1c}| + |m_{2c}^a - m_{2c}| = \mathrm{O}(n^{-1}) \tag{G.4}$$

uniformly in $z \in S_{out}(c_0, C_0)$. We postpone the proof of (G.4) until we complete the proof of Claim G.1. Now using $s_1^a \leqslant s_1$ we find that (A.16) holds for $m_{1,2c}^a$ and σ_1, s_1^a. Thus by Theorem A.7, we have that with high probability in Ω,

$$\max_{u,v\in\mathscr{A}} |\langle u, G_a(X,z)v\rangle - \langle u, \Pi_a(z)v\rangle| = \mathrm{o}(1)$$

uniformly in $z \in S_{out}(c_0, C_0)$. Hence to show (G.3), it suffices to prove that

$$\max_{1\leqslant i,j\leqslant k} |\langle Sw_i, [\Pi_a(z) - \Pi(z)]\, Sw_j\rangle| = \mathrm{o}(1). \tag{G.5}$$

Using (G.1), (G.2) and (G.4), we obtain that

$$\begin{aligned}
&\frac{-z}{m_{1c}(z)} w_i^\top S^\top [\Pi_a(z) - \Pi(z)]\, Sw_j \\
&= w_i^\top S^\top \left[\frac{1}{1 + m_{1c}(1-ee^\top)B(1-ee^\top)} \right. \\
&\quad \left. \times \left(Bee^\top + ee^\top B + ee^\top Bee^\top\right) \frac{1}{1+m_{1c}B} \right] Sw_j \\
&= \mathrm{o}(1) + w_i^\top S^\top \left[\frac{1}{1 + m_{1c}(1-ee^\top)B(1-ee^\top)} \right. \\
&\qquad \left. \times \left(ee^\top B\right) \frac{1}{1+m_{1c}B} \right] Sw_j \\
&= \mathrm{o}(1) + w_i^\top S^\top \left[\frac{1}{1 + m_{1c}(1-ee^\top)B(1-ee^\top)} \right. \\
&\qquad \left. \times m_{1c}^{-1} ee^\top \left(1 - \frac{1}{1+m_{1c}B}\right) \right] Sw_j \\
&= \mathrm{o}(1).
\end{aligned}$$

This concludes (G.5). □

Proof of (G.4). We claim that approximately, $m_{1,2c}$ satisfy the self-consistent equations for $m_{1,2c}^a$:

$$\begin{aligned}
m_{1c}(z) &= \frac{1}{n}\sum_{i=1}^p \frac{\sigma_i}{-z(1+\sigma_i m_{2c})}, \\
m_{2c}(z) &= \frac{1}{n}\sum_{\mu=1}^r \frac{s_\mu^a}{-z(1+s_\mu^a m_{1c})} + \mathrm{O}(n^{-1}).
\end{aligned} \tag{G.6}$$

Then (G.4) follows from Lemma 5.11 of [106], which gives the stability of the self-consistent equations. Roughly speaking, stability means that if (u_1, u_2) are satisfy the self-consistent equation for (m_{1c}^a, m_{2c}^a) up to some sufficiently small error ε, then we also have

$$|u_1(z) - m_{1c}(z)| + |u_2(z) - m_{2c}(z)| \lesssim \varepsilon$$

uniformly in $z \in S_{out}(c_0, C_0)$.

It remains to prove (G.6). The first equation is the first one in (A.12), while for the second equation, we claim that

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{\mu=1}^r \frac{s_\mu}{-z(1+s_\mu m_{1c})} - \frac{1}{n}\sum_{\mu=1}^r \frac{s_\mu^a}{-z(1+s_\mu^a m_{1c})} \right| \\
&= \mathrm{O}(n^{-1}).
\end{aligned} \tag{G.7}$$

For the imaginary part, we have

$$\begin{aligned}
&\mathrm{Im}\left(\frac{1}{n}\sum_{\mu=1}^r \frac{s_\mu}{1+s_\mu m_{1c}} - \frac{1}{n}\sum_{\mu=1}^r \frac{s_\mu^a}{1+s_\mu^a m_{1c}} \right) \\
&= -\frac{\mathrm{Im}\, m_{1c}}{n} \left(\sum_{\mu=1}^r \frac{1}{|s_\mu^{-1} + m_{1c}|^2} - \sum_{\mu=1}^r \frac{1}{|(s_\mu^a)^{-1} + m_{1c}|^2} \right).
\end{aligned}$$

By the Stieltjes transform

$$m_{1c}(z) = \int_0^{\lambda_+} \frac{\mathrm{d}\rho_{1c}(x)}{x - z},$$

we obtain that for $z = E + \mathrm{i}\eta \in S_{out}(c_0, C_0)$, $\mathrm{Im}\, m_{1c}(z) \geqslant 0$ and

$$0 > \mathrm{Re}\, m_{1c}(E) \geqslant m_{1c}(\lambda_+) \geqslant -s_1^{-1} \geqslant -(s_1^a)^{-1}.$$

Hence the function $|x + m_{1c}|^2$ is increasing in $x \in (s_1^{-1}, \infty)$. Using the Cauchy interlacing theorem

$$s_r^a \leqslant s_r \leqslant \cdots \leqslant s_2^a \leqslant s_2 \leqslant s_1^a \leqslant s_1,$$

we find that

$$\sum_{\mu=1}^r \frac{1}{|s_\mu^{-1} + m_{1c}|^2} - \sum_{\mu=1}^r \frac{1}{|(s_\mu^a)^{-1} + m_{1c}|^2} = \mathrm{O}(1).$$

Hence we obtain that

$$\mathrm{Im}\left(\frac{1}{n}\sum_{\mu=1}^r \frac{s_\mu}{1+s_\mu m_{1c}} - \frac{1}{n}\sum_{\mu=1}^r \frac{s_\mu^a}{1+s_\mu^a m_{1c}} \right) = \mathrm{O}(n^{-1}).$$

Together with a similar bound for the real part, we find (G.7), which concludes (G.6). □

Finally, we show that (G.1) and (G.2) hold (at least in probability) for all sketching methods we used in this paper.

(1) **Uniform random projection in Section III-A:** In this case $SS^\top = I_r$, so we only need to check that (G.1) holds. By the rotational invariance of S, we have

$$e^\top Sw_i \stackrel{d}{=} \frac{1}{\sqrt{r}} \sum_{i=1}^r \widehat{S}_{i1},$$

where $\widehat{S}$ is also an $r \times n$ uniform random projection matrix. By exchangeability, we have

$$\frac{1}{r}\mathbb{E}\Big| \sum_{i=1}^r \widehat{S}_{i1} \Big|^2 = \frac{1}{r}\mathbb{E} \sum_{i,j=1}^r \widehat{S}_{i1}\widehat{S}_{j1}$$

$$= \frac{1}{nr}\mathbb{E}\sum_{i,j=1}^{r}\sum_{\mu=1}^{n}\widehat{S}_{i\mu}\widehat{S}_{j\mu} = \frac{1}{n}.$$

Hence we have $e^\top S w_i \to 0$ in probability.

(2) **iid projection in Section III-B:** Note that $x_k := \sum_{l=1}^n S_{kl}w_i(l)$ are iid random variables with mean zero and variance n^{-1}. Hence by the law of large numbers (LLN), we have

$$e^\top S w_i = \frac{1}{\sqrt{r}}\sum_{k=1}^{r} x_k \to 0 \quad \text{a.s.}$$

For the bound (G.2), we use the local law, Theorem A.7. If we take $Y = S$, then

$$\frac{1}{m_{1c}(z)^{-1} + SS^\top}S$$

is (proportional to) the lower left block of $G(-m_{1c}(z)^{-1})$ in (A.21), and the local law (A.25) gives that (G.2) holds with high probability. If one is worried about $-m_{1c}(z)^{-1}$ may not be in the domain given in Theorem A.7, we remark that a local law in [12] for the $Y = S$ case was proved on a more general domain.

(3) **Random sampling in Section III-C:** In this case, the leading principal components of the centered model $\widetilde{\mathcal{Q}}_1^a = \widetilde{Y}_a^\top \widetilde{Y}_a$ will behave differently from those of the model $\widetilde{\mathcal{Q}}_1 = \widetilde{Y}\widetilde{Y}^\top$ under the sketching method (III.23). However, we can consider a slightly different random sampling $\widetilde{S}$ with random signs:

$$\widetilde{S}_{ii} = \varepsilon_i a_i, \tag{G.8}$$

where a_i is a Rademacher random variable uniform on $\{-1, 1\}$ independent of ε_i. In this case, we have that $Y^\top \widetilde{S}^\top \widetilde{S} Y = Y^\top S^\top S Y$, hence Theorem III.3 still holds for $\widetilde{\mathcal{Q}}_1$ under the sketching (G.8). On the other hand, we now check that (G.1) and (G.2) hold for $\widetilde{S}$, so that Theorem III.3 also holds for the centered model $\widetilde{\mathcal{Q}}_1^a$ under the sketching (G.8). Note that we have

$$\frac{1}{1 + m_{1c}(z)\widetilde{S}\widetilde{S}^\top}\widetilde{S} = \frac{1}{1 + m_{1c}(z)}\widetilde{S}. \tag{G.9}$$

Hence again we only need to check that (G.1) holds. We can calculate that

$$\mathbb{E}|e^\top \widetilde{S} w_i|^2 = \frac{1}{n}\mathbb{E}\Big|\sum_{l=1}^{n}\widetilde{S}_{ll}w_i(l)\Big|^2 = \frac{1}{n}\frac{r}{n}\sum_{l=1}^{n}|w_i(l)|^2 \leqslant \frac{1}{n}.$$

Hence we have $e^\top \widetilde{S} w_i \to 0$ in probability.

(4) **Randomized Hadamard sampling in Section III-D:** In this case, we also have (G.9) and hence we only need to check that (G.1) holds. We calculate that

$$\mathbb{E}|e^\top S w_i|^2 = \mathbb{E}|e^\top B_r z_i|^2 = \frac{1}{n}\mathbb{E}\Big|\sum_{l=1}^{n}(B_r)_{ll}z_i(l)\Big|^2.$$

For $l \neq l'$, we have

$$\mathbb{E}(z_i(l)z_i(l')) = \sum_{j,j'}\mathbb{E}[a_j^{(l)}a_j^{(l')}]w_i(j)w_i(j') = 0,$$

which gives that

$$\mathbb{E}|e^\top S w_i|^2 = \frac{1}{n}\frac{r}{n}\sum_{l=1}^{n}\mathbb{E}|z_i(l)|^2 \leqslant \frac{1}{n}.$$

Hence we have $e^\top S w_i \to 0$ in probability.

(5) **CountSketch in Section III-E:** In this case, we also have

$$\frac{1}{1 + m_{1c}(z)\widehat{S}\widehat{S}^\top}\widehat{S} = \frac{1}{1 + m_{1c}(z)}\widehat{S}$$

and hence we only need to check that (G.1) holds. Again we calculate the second moment of

$$e^\top \widehat{S} w_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{\mu:h(\mu)=i}\widehat{S}_{i\mu}w_i(\mu).$$

For $i \neq i'$, we have that

$$\mathbb{E}\sum_{\mu,\mu':h(\mu)=i,h(\mu')=i'}\widehat{S}_{i\mu}\widehat{S}_{i'\mu'}w_i(\mu)w_{i'}(\mu') = 0,$$

which leads to

$$\begin{aligned}\mathbb{E}|e^\top \widehat{S} w_i|^2 &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\sum_{\mu,\mu':h(\mu)=h(\mu')=i}\widehat{S}_{i\mu}\widehat{S}_{i\mu'}w_i(\mu)w_i(\mu') \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1(c_i \neq 0)}{c_i}\mathbb{E}\sum_{\mu:h(\mu)=i}w_i(\mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1(c_i \neq 0)}{c_i}\frac{c_i}{n}\sum_{\mu}w_i(\mu)^2 \leqslant \frac{1}{n},\end{aligned}$$

where we used the exchangeability in the third step. Hence we have $e^\top S w_i \to 0$ in probability.

ACKNOWLEDGMENT

We are grateful to the associate editor and two anonymous referees, whose feedback has resulted in a significant improvement. ED thanks Matan Gavish for inspiring discussions on this topic in 2014. We thank Ricardo Guerrero for valuable discussions.

REFERENCES

[1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

[2] G. W. Anderson and B. Farrell. Asymptotically liberating sequences of random unitary matrices. *Advances in Mathematics*, 255:381–413, 2014.

[3] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

[4] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, 2nd edition, 2010.

[5] Z. D. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 2 of *Mathematics Monograph Series*. Science Press, Beijing, 2006.

[6] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005.

[7] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

[8] Z. Bao, X. Ding, and K. Wang. Singular vector and singular subspace distribution for the matrix denoising model. *The Annals of Statistics*, 49(1):370 – 392, 2021.

[9] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[10] F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120 – 135, 2012.

[11] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[12] A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.*, 19(33):1–53, 2014.

[13] A. Bloemendal, A. Knowles, H.-T. Yau, and J. Yin. On the principal components of sample covariance matrices. *Prob. Theor. Rel. Fields*, 164(1):459–552, 2016.

[14] P. Bourgade, H.-T. Yau, and J. Yin. Local circular law for random matrices. *Probab. Theory Relat. Fields*, 159:545–595, 2014.

[15] C. J. Burges. Dimension reduction: A guided tour. *Foundations and Trends® in Machine Learning*, 2(4):275–365, 2010.

[16] H. M. Cann, C. De Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

[17] T. I. Cannings and R. J. Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

[18] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[19] S. Chatterjee. A generalization of the lindeberg principle. *The Annals of Probability*, 34(6):2061–2076, 2006.

[20] S. Chen, Y. Liu, M. R. Lyu, I. King, and S. Zhang. Fast relative-error approximation algorithm for ridge regression. In *UAI*, pages 201–210, 2015.

[21] S. Chen, R. Varma, A. Singh, and J. Kovačcević. A statistical perspective of sampling scores for linear regression. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 1556–1560. IEEE, 2016.

[22] K. L. Clarkson and D. P. Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):54, 2017.

[23] L. Cordero-Grande. Mixandmix: numerical techniques for the computation of empirical spectral distributions of population mixtures. *Computational Statistics & Data Analysis*, 141:1–11, 2020.

[24] R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[25] R. Couillet, M. Debbah, and J. W. Silverstein. A deterministic equivalent for the analysis of correlated mimo multiple access channels. *IEEE Trans. Inform. Theory*, 57(6):3493–3514, 2011.

[26] R. Couillet and W. Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 03(04):1450016, 2014.

[27] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

[28] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388, 2015.

[29] X. Ding. High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli*, 26(1):387–417, 2020.

[30] X. Ding and F. Yang. Spiked separable covariance matrices and principal components. *The Annals of Statistics*, 49(2):1113 – 1138, 2021.

[31] E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.

[32] E. Dobriban. Eigenedge. https://github.com/dobriban/eigenedge, 2015.

[33] E. Dobriban. Sharp detection in PCA under correlations: all eigenvalues matter. *The Annals of Statistics*, 45(4):1810–1833, 2017.

[34] E. Dobriban. Permutation methods for factor analysis and PCA. *The Annals of Statistics*, 48(5):2824 – 2847, 2020.

[35] E. Dobriban and S. Liu. Asymptotic for sketching in least squares regression. *arXiv preprint arXiv:1810.06089, NeurIPS 2019*, 2019.

[36] E. Dobriban and A. B. Owen. Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):163–183, 2019.

[37] E. Dobriban and Y. Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918 – 943, 2021.

[38] D. Donoho, M. Gavish, and I. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.*, 46(4):1742–1778, 08 2018.

[39] D. L. Donoho and B. Ghorbani. Optimal covariance estimation for condition number loss in the spiked model. *arXiv preprint arXiv:1810.07403*, 2018.

[40] P. Drineas and M. W. Mahoney. Lectures on randomized numerical linear algebra. *arXiv preprint arXiv:1712.08880*, 2017.

[41] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

[42] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

[43] N. El Karoui. Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6):2362–2405, 2009.

[44] L. Erdős, A. Knowles, and H.-T. Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14:1837–1926, 2013.

[45] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Delocalization and diffusion profile for random band matrices. *Commun. Math. Phys.*, 323:367–416, 2013.

[46] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. The local semicircle law for a general class of random matrices. *Electron. J. Probab.*, 18:1–58, 2013.

[47] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin. Spectral statistics of Erdős-Rényi graphs I: Local semicircle law. *Ann. Probab.*, 41(3B):2279–2375, 05 2013.

[48] L. Erdős and H.-T. Yau. Universality of local spectral statistics of random matrices. *Bulletin of the American Mathematical Society*, 49(3):377–414, 2012.

[49] L. Erdos and H.-T. Yau. A dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics*, 28, 2017.

[50] N. B. Erichson, S. Voronin, S. L. Brunton, and J. N. Kutz. Randomized matrix decompositions using R. *Journal of Statistical Software, Articles*, 89(11):1–48, 2019.

[51] J. Fan, Q. Sun, W.-X. Zhou, and Z. Zhu. Principal component analysis for big data. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2014.

[52] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

[53] K. J. Galinsky, G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, and A. L. Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

[54] 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.

[55] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

[56] M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

[57] V. Girko. Random matrices. *Handbook of Algebra, ed. Hazewinkel*, 1:27–78, 1975.

[58] V. L. Girko. Spectral theory of random matrices. *Russian Mathematical Surveys*, 40(1):77, 1985.

[59] V. L. Girko. *Theory of random determinants*, volume 45. Springer Science & Business Media, 2012.

[60] W. Hachem, P. Loubaton, and J. Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

[61] N. Halko, P.-G. Martinsson, Y. Shkolnisky, and M. Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific computing*, 33(5):2580–2594, 2011.

[62] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[63] D. Homrighausen and D. J. McDonald. On the nyström and column-sampling methods for the approximate principal components analysis of large datasets. *Journal of Computational and Graphical Statistics*, 25(2):344–362, 2016.
[64] T. Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *The Annals of Probability*, 34(4):1497–1529, 2006.
[65] T. Jiang and Y. Ma. Distances between random orthogonal matrices and independent normals. *Transactions of the American Mathematical Society*, 372(3):1509–1553, 2019.
[66] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 2001.
[67] I. M. Johnstone and D. Paul. Pca in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.
[68] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.
[69] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):1–23, 2014.
[70] A. Knowles and J. Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, pages 1–96, 2016.
[71] J. Lacotte, S. Liu, E. Dobriban, and M. Pilanci. Optimal iterative sketching methods with the subsampled randomized hadamard transform. *Advances in Neural Information Processing Systems*, 33, 2020.
[72] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, 2008.
[73] E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.
[74] S. Liu and E. Dobriban. Ridge regression: Structure, cross-validation, and sketching. *arXiv preprint arXiv:1910.02373, International Conference on Learning Representations (ICLR) 2020*, 2019.
[75] M. Lopes, L. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.
[76] Y. Lu, P. Dhillon, D. P. Foster, and L. Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems*, pages 369–377, 2013.
[77] P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, 16(1):861–911, 2015.
[78] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.
[79] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.
[80] C. Musco and C. Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. *arXiv preprint arXiv:1504.05477, NeurIPS 2015*, 2015.
[81] R. R. Nadakuditi. Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60(5):3002–3018, 2014.
[82] J. Nelson and H. L. Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 117–126, 2013.
[83] S. Ng. Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Technical report, National Bureau of Economic Research, 2017.
[84] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.
[85] D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.
[86] M. Pilanci and M. J. Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.
[87] M. Pilanci and M. J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879, 2016.
[88] M. Pilanci and M. J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.
[89] G. Raskutti and M. W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538, 2016.
[90] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.
[91] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.
[92] M. J. Schneider and S. Gupta. Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, 32(2):243–256, 2016.
[93] J. W. Silverstein and S.-I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.*, 54(2):295–309, 1995.
[94] C. Spearman. "General Intelligence", objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.
[95] R. Srivastava, P. Li, and D. Ruppert. Rapt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970, 2016.
[96] G. W. Stewart. *Matrix perturbation theory*. Citeseer, 1990.
[97] L. L. Thurstone. *Multiple-factor analysis*. University of Chicago Press, 1947.
[98] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
[99] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1454–1485, 2017.
[100] S. S. Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.
[101] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
[102] S. Wang, A. Gittens, and M. W. Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18:1–50, 2018.
[103] H. Wickham. *nycflights13: Flights that Departed NYC in 2013*, 2018. R package version 1.0.0.
[104] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.
[105] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.
[106] F. Yang. Edge universality of separable covariance matrices. *Electron. J. Probab.*, 24:57 pp., 2019.
[107] J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, New York, 2015.
[108] L. Zhang. *Spectral Analysis of Large Dimensional Random Matrices*. PhD thesis, National University of Singapore, 2006.

Fan Yang received the Ph.D. degree in mathematics from the University of California, Los Angeles in 2019 and the Ph.D. degree in physics from the Chinese University of Hong Kong in 2014. He is currently a postdoctoral researcher with the Wharton Statistics Department, University of Pennsylvania. His research interests include random matrix theory, high-dimensional statistics, statistical machine learning and mathematical physics.

Sifan Liu is currently pursuing the Ph.D. degree in the Statistics Department at Stanford University. Her research interests include Monte Carlo and quasi-Monte Carlo methods, selective inference and scalable algorithms.

Edgar Dobriban is an assistant professor of Statistics and Data Science, with secondary appointment in Computer and Information Science, at the University of Pennsylvania. Prior to that, he received a PhD degree from Stanford University in 2017. His research interests include the statistical analysis of large-scale data, and the theoretical foundations of machine learning.

David P. Woodruff is an associate professor at Carnegie Mellon University in the Computer Science Department, where he has been since 2017. Prior to that, he was in the Algorithms and Complexity group at IBM Almaden for ten years, and prior to that he received his PhD from MIT. His interests are in algorithms, machine learning, and optimization, with an emphasis on techniques for dimensionality reduction.