Sample-Measurement Tradeoff in Support Recovery under a Subgaussian Prior

Lekshmi Ramesh, Chandra R. Murthy, and Himanshu Tyagi

Abstract

Data samples from $\mathbb{R}^d$ with a common support of size k are accessed through m random linear projections per sample. It is well-known that roughly k measurements from a single sample are sufficient to recover the support. In the multiple sample setting, do k overall measurements still suffice when only $m < k$ measurements per sample are allowed? We answer this question in the negative by considering a generative model setting with independent samples drawn from a subgaussian prior. We show that $n = \Theta((k^2/m^2) \cdot \log k(d-k))$ samples are necessary and sufficient to recover the support exactly. In turn, this shows that k overall samples are insufficient for support recovery when $m < k$; instead we need about k^2/m overall measurements. Our proposed sample-optimal estimator has a closed-form expression, has computational complexity of $O(dnm)$, and is of independent interest.

I. Introduction

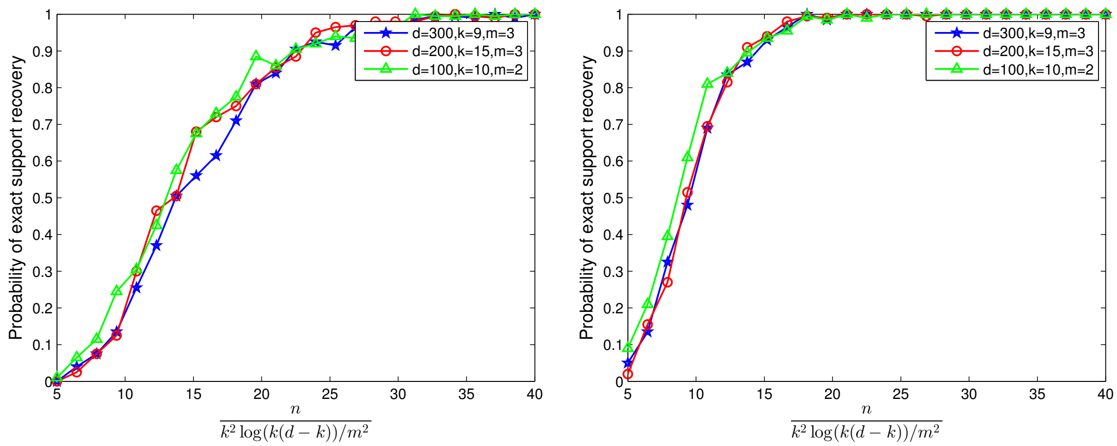
A set of n vectors has a common support of cardinality k that is much smaller than the dimension d of the vectors. It is easy to find this common support by simply looking at a single vector. But this will require d measurements, one for checking each coordinate of the vector. As is now well-known, we can make do with $m = O(k \log(d-k))$ random linear measurements on a single vector to recover the support [31]. When multiple samples are available, one can try to estimate the support by considering each sample in isolation, but it requires $m > k$ measurements and ignores the fact that the samples share a common support. A natural question that arises then is whether we can still recover the unknown support with k *overall* measurements (i.e., would $nm \leq k$ suffice)? We examine this question in a natural Bayesian setting and answer it in the negative: when $m < k$, we will need at least k^2/m overall measurements. Thus, in sharp contrast with the $m > k$ regime where k overall measurements suffice, a much larger number of overall measurements are necessary when $m < k$.

We start with the simpler Gaussian setting, and discuss the more general subgaussian setting in later sections. Specifically, consider independent d-dimensional samples $X_1, \ldots, X_n$ where each X_i is a zero-mean Gaussian vector with a diagonal covariance matrix $\mathtt{diag}(\lambda)$. We assume that the diagonal entry λ_i, which represents the variance along the ith coordinate, is either 0 or 1, whereby the common support of the vectors coincides with the locations of 1s. The assumption that λ is binary can be relaxed, as we discuss in section V. We make linear measurements on the vectors X_i using independent random Gaussian matrices Φ_i with columns that have unit expected squared norms. The goal is to recover the common support using measurements $Y_i = \Phi_i X_i$, $1 \leq i \leq n$. We show that in the *measurement-starved* regime of $m < \alpha k$ with $\alpha < 1$, the minimum number of samples required to recover the support correctly with large probability is $\Theta((k^2/m^2) \log k(d-k))$ (assuming $m \geq (\log k)^2$).

The sample-optimal estimator we propose entails forming an estimate $\tilde{\lambda}$ of λ and then obtaining the support by selecting the k largest entries of $\tilde{\lambda}$. The estimate $\tilde{\lambda}_i$ has a closed-form expression: it is simply the empirical average $\frac{1}{n}\sum_{j=1}^{n}(\Phi_{ji}^\top Y_j)^2$ where Φ_{ji} denotes the ith column of the jth measurement matrix. This is in contrast

The authors are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India. Email: {lekshmi, cmurthy, htyagi}@iisc.ac.in.

This work was financially supported by a research fellowship from the Ministry of Electronics and Information Technology, Govt. of India.



(a) Gaussian prior (b) Rademacher prior

Fig. 1: Phase transition of the closed-form estimator.

to the standard Sparse Bayesian Learning (SBL) approach where the maximum likelihood (ML) estimate of λ is used, which can only be expressed as a nonconvex optimization problem and does not have a closed-form solution. Furthermore, the proposed estimator works under a much broader setting, one with subgaussian prior on X_1^n and subgaussian measurement matrices, and with additive subgaussian noise. Note that under these minimal assumptions, the underlying statistical problem becomes nonparametric, and an ML estimate for λ cannot be expressed as a closed-form optimization problem. In Figure 1, we plot the probability of exact support recovery for our proposed closed-form estimator as a function of the normalized number of samples. These curves confirm our theoretical results, and a clear phase transition can be seen. We discuss these numerical evaluations further in Section III-C.

Our information-theoretic lower bound is obtained by applying Fano's method to a difficult case with supports of size k differing in one entry. The main challenge in the proof is to characterize the reduction in the distances between the distributions corresponding to different supports due to linear measurements. We capture this by a quantity related to the spectrum of the Gram matrix of the random Gaussian measurement matrix.

Before describing the related literature, we make a distinction between support recovery and recovery of the data vectors. When $m \geq k$, support recovery implies recovery of the data vectors also. Indeed, given the support, one can estimate the data vectors by solving a least squares problem restricted to the support. When $m < k$, although support recovery is still possible as we show in this work, data recovery is no longer possible, since there are infinitely many solutions even after restricting to the support. Recovery of the support, rather than the data vectors, is important in several practical applications including spectrum sensing [26] and group testing [33].

Information-theoretically optimal support recovery in the single sample setting is well-understood; [31], [12] and [2] were some of the first works to look at this problem. In particular, [31] shows that for a deterministic input vector, $m = \Theta(k \log(d-k))$ measurements are necessary and sufficient to exactly recover the support using a Gaussian measurement matrix, establishing that support recovery is impossible in the $m < k$ regime using a single sample. Following these works, several papers have extended results to the multiple sample setting. We would like to point out that the settings considered in these works vary in terms of whether the same measurement matrix is used for all samples or different measurement matrices are used across samples; whether the data vectors are deterministic or sampled from a generative model; and whether the measurements are noisy or noiseless. In our setting, as we mentioned, we consider $\{X_i\}_{i=1}^n$ generated from a certain generative model, random measurement matrices

$\{\Phi_i\}_i^n$ chosen independently across samples and independent of the data vectors, and measurements corrupted by subgaussian noise. We emphasize that both our upper and lower bounds are for random measurement matrix designs. There has also been work on approximate support recovery [24], [25] in the single sample setting. In this work, however, our focus will be on exact recovery.

A setup similar to ours was studied in [22], but the results are not tight in the $m < k$ regime. In particular, [22] showed a lower bound on sample complexity of support recovery of roughly (k/m), much weaker than our $(k/m)^2$ lower bound. Another related line of works [27], [17] studies this problem considering the *same* measurement matrix for all samples, under the assumption that the data vectors are deterministic. In [17], the authors connect the support recovery problem to communication over a single input multiple output MAC channel. However, the performance guarantees are asymptotic in nature ($d \to \infty$ and $k,\ n$ fixed). In [27], it is shown that using a Gaussian measurement matrix, the probability of error of the maximum likelihood decoder goes to zero with d provided $m \geq k \log \frac{d}{k}$ and $n \geq \frac{\log d}{\log \log d}$. Also, several algorithms from the single sample setting have been generalized to work with multiple samples that include convex programming methods [20], [28], [11], thresholding-based methods [13], [14], Bayesian methods [34] and greedy methods [29], [30]. However, none of the above works addresses the question of tradeoff between m and n when $m < k$. Initial works considering the $m < k$ regime were [21] and [4], followed by [18] and [23], where it was empirically demonstrated that when multiple samples are available, it is possible to operate in the $m < k$ regime. However, the analysis in [4] is done under two fairly restrictive conditions. The first condition is an orthogonality assumption on the data vectors that requires $\sum_{i=1}^n X_i X_i^\top$ to be diagonal. In fact, as we show in our analysis, a much weaker assumption is sufficient for randomly generated data vectors with independent coordinates, in which case the condition holds in expectation. The second condition is that $m^2 \geq d$, which is much stronger than our $m \geq (\log k)^2$ condition[1]

In [21], a LASSO-based approach is proposed to recover the common support using correlation among the $X_i s$. The authors empirically show that support recovery is possible using the same measurement matrix across samples (with large n) for support size $k \geq m$ and conjecture that k can be as large as $O(m^2)$. In another recent work closely related to ours [19], the authors demonstrate the possibility of operating in the $m < k$ regime. Their results are for the same measurement matrix across samples and for $\{X_i\}_{i=1}^n$ drawn from a certain prior. Similar to [27], the authors analyze the exhaustive search ML decoder and show that i¡ts error probability decays exponentially with the number of samples n. The error exponent, however, is expressed in terms of the eigenvalues of certain matrices and its exact dependence on the parameters k, m and d is not clear.

Our formulation of support recovery in a Bayesian setting naturally relates to some of the works on covariance estimation. A recent work which looks at the problem of covariance estimation from low-dimensional projections of the data is [3]. No structural assumptions are made on the covariance matrix. As we will see in the next section, the support recovery problem amounts to estimating a diagonal and sparse covariance matrix, and the general results from [3, Corollary 3] for this specific case are loose and do not give the correct scaling for sample complexity. Two other works that study covariance estimation from projected samples are [7] and [8], focusing specifically on the $m = 1$ case. In particular, [7, Theorem 4.1] assumes a low-rank plus identity structure (the spiked covariance model) on the data covariance matrix and shows that $n = O(rd)$ samples are sufficient for recovery via convex programming, where r is the rank of the true covariance matrix. Similar results are obtained in [8], which also provides results for Toeplitz-structured covariance matrices. However, a direct application of these results to the diagonal sparse case does not give the correct scaling on the number of samples. Also, since m is set to one, the tradeoff between m and n is not clear.

[1] This condition, too, can be avoided if we replace some union bounds in our proof with appropriate bounds for heavy tails obtained using a tail-splitting technique; we omit this more technical approach to keep our proofs conceptually simpler.

Our setting is also related to the recently considered inference under local information constraints setting of [1]. We impose information constraints on each sample by allowing only m linear measurements per sample. Roughly, our results say that when local information constraints are placed (namely, $m < k$), support recovery requires much more than k overall measurements.

Organization. In the next section, we formally state the problem and the assumptions we make in our generative model setting. We then state our main result, which characterizes the sample complexity of support recovery. In section III, we propose and analyze a closed-form estimator and show it to be sample-optimal. We present simulation results, followed by the proof of the lower bound in section IV. We conclude with discussion and future work in section VI. Some of the technical details have been relegated to the appendix. Appendices A and B contain the proof of the main technical result needed for analyzing our scheme. Appendix C contains a bound for the fourth moment of the minimum eigenvalue of a Wishart matrix, which is required for the lower bound and may be of independent interest. The background needed for our proofs is reviewed in Appendix D.

Notation and Preliminaries. For a matrix A_j, A_{ji} denotes its ith column, $A_j(u,v)$ denotes its (u,v)th entry and $(A_j)_S$ denotes the submatrix formed by columns indexed by S. To denote the set $\{A_j\}_{i=1}^n$ of matrices (or vectors), we use the shorthand A_1^n. Also, for a vector X_j, X_{ji} denotes the ith component of X_j. For symmetric matrices A and B, the notation $A \succcurlyeq B$ denotes that the matrix $A - B$ is positive semidefinite. For a vector $a \in \mathbb{R}^d$, $\texttt{diag}(a)$ denotes a diagonal matrix with the entries of a on the diagonal. A random variable X is subgaussian with variance parameter σ^2, denoted $X \sim \text{subG}(\sigma^2)$, if

$$\log \mathbb{E}\left[e^{\theta(X-\mathbb{E}[X])}\right] \leq \theta^2\sigma^2/2, \tag{1}$$

for all $\theta \in \mathbb{R}$. Similarly, a random variable X is subexponential with parameters v^2 and $b > 0$, denoted $X \sim \text{subexp}(v^2, b)$, if

$$\log \mathbb{E}\left[e^{\theta(X-\mathbb{E}[X])}\right] \leq \theta^2 v^2/2 \tag{2}$$

for all $|\theta| < 1/b$. When taking expectation of a function of several random variables $Z = f(X_1, \ldots, X_n)$, we use $\mathbb{E}_{X_1}[Z]$ to denote that the expectation is with respect to the distribution of X_1.

II. Problem formulation and main result

We start with the basic setting of Gaussian prior with noiseless measurements obtained using Gaussian sensing matrices. However, as we shall see later, our results generalize to much broader settings and extend to subgaussian priors on data and noisy subgaussian measurements.

In the basic setting, we consider a Bayesian formulation for support recovery where the input comprises n independent samples $X_1, \ldots, X_n$ in $\mathbb{R}^d$, with each X_i having a zero-mean Gaussian distribution. We denote the covariance of X_i by $K_\lambda \stackrel{\text{def}}{=} \texttt{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, where the d-dimensional vector λ has entries $\lambda_1, \lambda_2, \ldots, \lambda_d$ such that $\lambda \in \mathcal{S}_{k,d} \stackrel{\text{def}}{=} \left\{u \in \{0,1\}^d : \|u\|_0 = k\right\}$. That is, the (random) data vectors have a common support $S = \texttt{supp}(\lambda)$ of size k.

Each X_i is passed through a random $m \times d$ measurement matrix Φ_i, $1 \leq i \leq n$, with independent, zero-mean Gaussian entries with variance $1/m$, and the observations $Y_i = \Phi_i X_i \in \mathbb{R}^m$ are made available to a center. Using the measurements $Y_1, \ldots, Y_n$, the center seeks to determine the common support S of $X_1, \ldots, X_n$.

To that end, the center uses an estimate $\hat{S} : \mathbb{R}^{m \times n} \to \binom{[d]}{k}$, where $\binom{[d]}{k}$ denotes the set of all subsets of $[d]$ of cardinality k. We seek estimators that can recover the support of λ accurately with probability of error no more than $\delta \stackrel{\text{def}}{=} 1/3$,[2] namely

$$\Pr\left(\hat{S}(Y^n) \neq \texttt{supp}(\lambda)\right) \leq \delta, \quad \forall\, \lambda \in \mathcal{S}_{k,d}. \tag{3}$$

In the compressed sensing literature, this is usually referred to as a non-uniform recovery guarantee.

We are interested in sample-efficient estimators. The next definition introduces the fundamental quantity of interest to us.

Definition 1 (Sample complexity of support recovery)**.** For $m, k, d \in \mathbb{N}$, the sample complexity of support recovery $n^*(m, k, d)$ is defined as the minimum number of samples n for which we can find an estimator $\hat{S}$ satisfying (3).

Remark 1. Our formulation assumes that the support size k is known. That said, our proposed estimator extends easily to the setting where we only have an upper bound of k on the support size, and we seek to output a set of indices containing the support.

Our main result is the following.

Theorem 1 (Characterization of sample complexity)**.** *For* $(\log k)^2 \leq m < k/2$ *and* $1 \leq k \leq d-1$*, the sample complexity of support recovery in the setting above is given by*

$$n^*(m, k, d) = \Theta\left(\frac{k^2}{m^2} \cdot \log k(d-k)\right).$$

Remark 2. We expect the scaling in Theorem 1 to hold good even when $m < (\log k)^2$. In fact, our lower bound result continues to hold for $m = 1$. The current upper bound proof, however, requires $m \geq (\log k)^2$. We comment on this more in Appendix A.

We provide the optimal estimator and prove the upper bound in Section III and the information-theoretic lower bound in Section IV. Our proof yields a lower bound when $m < \alpha k$ for any $\alpha < 1$.

Our proposed estimator and its analysis applies to a much broader setting involving subgaussian priors (see (1) for definition). For X_1^n, we can use any prior with subgaussian distributed entries, i.e., the entries of X_i are independent and zero-mean with $\mathbb{E}\left[X_{i,j}^2\right] = \lambda_j$ for $\lambda \in \mathcal{S}_{k,d}$ and $X_{i,j} \sim \text{subG}(\lambda'_j)$, where λ'_j is the variance parameter for the subgaussian random variable $X_{i,j}$. Our analysis will go through as long as the variance and variance parameters differ only up to a constant factor.

Also, the measurement matrices Φ_i can be chosen to have independent, zero-mean subgaussian distributed entries in place of Gaussian. However, as above, we assume that the variance and variance parameter of each entry are the same up to a multiplicative constant factor. Further, we assume that the fourth moment of the entries of Φ_i is of the order of the square of the variance. Two important ensembles that satisfy these properties are the Gaussian ensemble and the Rademacher ensemble.

For clarity, we restate our assumptions below. These assumptions are required for the analysis of our estimator; the lower bound proof is done under the more restrictive setting of Gaussian measurement matrix ensemble (which implies a lower bound for the subgaussian ensemble also).

Assumption 1. *The entries of* X_i, $i \in [n]$, *are independent and zero-mean with* $\mathbb{E}\left[X_{i,j}^2\right] = \lambda_j$ *for* $\lambda \in \mathcal{S}_{k,d}$ *and* $X_{i,j} \sim$ *subG*$(c\lambda_j)$, *where* c *is an absolute constant;*

Assumption 2. *The entries of* Φ_i, $i \in [n]$, *are independent and zero-mean with* $\mathbb{E}\left[\Phi_i(u,v)^2\right] = 1/m$, $\Phi_i(u,v) \sim$ *subG*(c'/m), *and* $\mathbb{E}\left[\Phi_i(u,v)^4\right] = c''/m^2$, *where* c' *and* c'' *are absolute constants.*

[2]Note that the value $\delta = 1/3$ is chosen here for convenience and can be replaced with any acceptable value below $1/2$. However, our results may not be tight in their dependence on δ.

Remark 3. Our results also extend to the case when the data vectors are not necessarily sparse in the standard basis for $\mathbb{R}^d$, i.e., the data vectors can be expressed as $X_i = BZ_i$, $i \in [n]$, where B is any orthonormal basis for $\mathbb{R}^d$ and Z_is have a common support of size k. Under the same generative model as before, but for Z_is this time, Theorem 1 continues to hold. This is because when Φ_i is Gaussian, the *effective* measurement matrix $\Phi_i B$ also satisfies the properties we mentioned above, namely, it has independent mean zero Gaussian entries with variance $1/m$ and fourth moment $3/m^2$.

Remark 4. We have restricted λ to binary vectors for the ease of presentation. Later, in Section V, we will show that our sample complexity results extend almost verbatim to a more general setting with the nonzero coordinates of λ taking values between $\lambda_{\min}$ and $\lambda_{\max}$. The only change, in effect, is a factor $(\lambda_{\max}/\lambda_{\min})^2$ blow-up in the sample complexity of support recovery.

Finally, we can allow noisy measurements $Y_i = \Phi_i X_i + W_i \in \mathbb{R}^m$ where the noise W_i has independent, zero-mean subgaussian entries independent of X_i and Φ_i, with variance parameter σ^2.

We present the upper bound for this more general setting, along with our proposed estimator, in the next section.

III. The estimator and its analysis

We will work with the more general setting described in the previous section, that is with subgaussian random variables satisfying assumptions 1 and 2. In fact, for simplicity, we assume that $X_{i,j}$ and W_i are subgaussian with variance parameter equal to their respective variances, a property known as *strict subgaussianity*. Also, for the measurement matrix, we work with the same parameters as those for the Gaussian ensemble and set

$$\mathbb{E}\left[\Phi_i(u,v)^2\right] = \frac{1}{m}, \qquad \mathbb{E}\left[\Phi_i(u,v)^4\right] = \frac{3}{m^2},$$

and assume that $\Phi_i(u,v)$ is subgaussian with variance parameter $1/m$. These assumptions of equality can be relaxed to order equality up to multiplicative constants.

A. *The estimator*

We now present our closed-form estimator for λ. To build heuristics, consider the trivial case where we can directly access samples $\{X_i\}_{i=1}^n$. Then, a natural estimate for λ_i is the sample variance. But in our setting, we only have access to the measurements $\{Y_i\}_{i=1}^n$. We compute the vector $\Phi_i^\top Y_i$ and treat it as a "proxy" for X_i. When $\Phi_i^\top \Phi_i = I$ and the measurements are noiseless, this proxy will indeed coincide with X_i. We compute the sample variances using these new proxy samples and use it to find an estimate for the support of λ.

Formally, we consider the estimate A for the covariance matrix of X_is given by

$$A = \frac{1}{n}\sum_{j=1}^{n} \Phi_j^\top Y_j Y_j^\top \Phi_j.$$

Note that A is positive semidefinite. We form an intermediate estimate $\tilde{\lambda}$ for the variance vector λ using the diagonal entries of A as follows:

$$\begin{aligned}\tilde{\lambda}_i \stackrel{\text{def}}{=} A_{ii} &= \frac{1}{n}\sum_{j=1}^{n} (\Phi_j^\top Y_j Y_j^\top \Phi_j)_{ii} \\ &= \frac{1}{n}\sum_{j=1}^{n} (\Phi_{ji}^\top Y_j)^2, \end{aligned} \tag{4}$$

where Φ_{ji} denotes the ith column of Φ_j. Since we are only interested in estimating the support, we simply declare indices corresponding to the largest k entries of $\tilde{\lambda}$ as the support, namely, we sort $\tilde{\lambda}$ to get $\tilde{\lambda}_{(1)} \geq \tilde{\lambda}_{(2)} \geq \cdots \geq \tilde{\lambda}_{(d)}$ and output

$$\tilde{S} = \{(1), ..., (k)\}, \tag{5}$$

where (i) denotes the index of the ith largest entry in $\tilde{\lambda}$. This is similar in spirit to the Iterative Hard Thresholding (IHT) algorithm [5] from the compressed sensing literature, where a similar support estimation step followed by least squares is used to estimate the data vectors. The difference is that IHT is an iterative procedure and the least squares step requires $m \geq k$. Also note that evaluating $\tilde{\lambda}_i$ requires $O(nm)$ steps, whereby the overall computational complexity of (naively) evaluating our proposed estimator is $O(dnm)$.

Before we move to detailed analysis in the next section, we do a quick sanity test for our estimator and evaluate its "expected behavior". An easy calculation shows that $\tilde{\lambda}_i$ is an estimate of λ_i with a constant bias depending only on k, m, and σ^2. In particular, we have the following result.

Lemma 1. *Let the estimator $\tilde{\lambda}$ be as defined in (4). Then, under Assumptions 1 and 2 with $c = c' = c'' = 1$, we have that*

$$\mathbb{E}\left[\tilde{\lambda}_i\right] = \frac{m+1}{m}\lambda_i + \frac{k}{m} + \sigma^2, \quad i \in [d],$$

where the expectation is with respect to the joint distribution of $\{X_1^n, \Phi_1^n, W_1^n\}$.

Proof. Recall that $\tilde{\lambda}_i = \frac{1}{n}\sum_{j=1}^n (\Phi_{ji}^\top Y_j)^2$ for $i \in [d]$. Since $Y_j = \Phi_j X_j + W_j$, we can rewrite the estimator as

$$\tilde{\lambda}_i = \frac{1}{n}\sum_{j=1}^n \left(\sum_{l\in S} X_{jl}(\Phi_{ji}^\top \Phi_{jl}) + \Phi_{ji}^\top W_j\right)^2, \quad i \in [d].$$

Taking expectation, we note that for $i \in S$,

$$\begin{aligned}
\mathbb{E}\left[\tilde{\lambda}_i\right] &= \mathbb{E}_{\Phi_1^n}\left[\mathbb{E}_{X_1^n, W_1^n}\left[\frac{1}{n}\sum_{j=1}^n \left(\sum_{l\in S} X_{jl}(\Phi_{ji}^\top \Phi_{jl}) + \Phi_{ji}^\top W_j\right)^2 \middle| \Phi_1^n\right]\right] \\
&= \mathbb{E}_{\Phi_1^n}\left[\frac{1}{n}\sum_{j=1}^n \left(\mathbb{E}_{X_1^n}\left[\sum_{l\in S} X_{jl}^2(\Phi_{ji}^\top \Phi_{jl})^2\right] + \mathbb{E}_{W_1^n}\left[(\Phi_{ji}^\top W_j)^2\right]\right)\middle| \Phi_1^n\right] \\
&= \mathbb{E}_{\Phi_1^n}\left[\frac{1}{n}\sum_{j=1}^n \left(\|\Phi_{ji}\|_2^4 + \sum_{l\in S\setminus\{i\}} (\Phi_{ji}^\top \Phi_{jl})^2 + \sigma^2\|\Phi_{ji}\|_2^2\right)\right],
\end{aligned}$$

where the second step uses the fact that X_j has zero mean entries. A similar calculation shows that for $i \in S^c$,

$$\mathbb{E}\left[\tilde{\lambda}_i\right] = \mathbb{E}_{\Phi_1^n}\left[\frac{1}{n}\sum_{j=1}^n \left(\sum_{l\in S} (\Phi_{ji}^\top \Phi_{jl})^2 + \sigma^2\|\Phi_{ji}\|_2^2\right)\right].$$

Using our assumption that the columns of Φ_j have independent mean-zero entries with variance $1/m$ and fourth moment $3/m^2$, it follows from Lemma D.4 that for $i \in S$,

$$\mathbb{E}\left[\tilde{\lambda}_i\right] = 1 + \frac{k+1}{m} + \sigma^2,$$

and for $i \in S^c$,

$$\mathbb{E}\left[\tilde{\lambda}_i\right] = \frac{k}{m} + \sigma^2.$$

Combining the two results above, we get

$$\mathbb{E}\left[\tilde{\lambda}_i\right] = \frac{m+1}{m}\lambda_i + \frac{k}{m} + \sigma^2, \quad i \in [d],$$

which establishes the lemma. □

We work with this biased $\tilde{\lambda}$ and analyze its performance in the next section. Since the bias is the same across all coordinates, it does not affect sorting/thresholding based procedures. The key observation here is that the expected values of the coordinates of $\tilde{\lambda}$ in the support of λ exceeds those outside the support, making it an appropriate statistic for support recovery.

B. And its analysis

A high level overview of our analysis is as follows. We first note that, conditioned on the measurement matrices, the entries of $\tilde{\lambda}$ are sums of independent, subexponential random variables (defined in (2)). If we can ensure that there is sufficient separation between the typical values of $\tilde{\lambda}_i$ in the $i \in S$ and $i' \in S^c$ cases, then we can distinguish between the two cases. We show that such a separation holds with high probability for our subgaussian measurement matrix ensemble.

We now present the performance of our estimator.

Theorem 2. *Let $\tilde{S}$ be the estimator described in (5) and assume that $(\log k)^2 \leq m$. Then, under Assumptions 1 and 2, $\tilde{S}$ equals the true support with probability at least $1-\delta$ provided*

$$n \geq c\left(\frac{k}{m} + 1 + \sigma^2\right)^2 \log\frac{k(d-k)}{\delta}, \tag{6}$$

for an absolute constant c.

Remark 5. We note that the result above applies for all k and all $m > (\log k)^2$, and not only to our regime of interest $m < k$. When $\sigma^2 = 0$ and $k/m > 1$, we obtain the upper bound claimed in Theorem 1.

Remark 6. Our lower bound on sample complexity is for the noiseless case and therefore does not capture the dependence on σ^2. However, simulation results (see Figure 2 in Section III-C) suggest that the dependence in (6) is tight.

Proof. While computationally tractable, analyzing our proposed estimator directly may not be easy. Instead, we analyze an alternative thresholding-based estimator given by

$$\hat{\lambda}_i = \mathbb{1}_{\{\tilde{\lambda}_i \geq \tau\}}. \tag{7}$$

We note that if $\lambda = \hat{\lambda}$, the largest k entries of $\tilde{\lambda}$ must coincide with the support of λ. Therefore,

$$\Pr\left(\tilde{S} \neq \mathtt{supp}(\lambda)\right) \leq \Pr\left(\hat{S} \neq \mathtt{supp}(\lambda)\right), \tag{8}$$

where $\hat{S}$ is the support of $\hat{\lambda}$. Using this observation, it suffices to analyze the estimator $\hat{\lambda}$ in (7), which will be our focus below.

The proof of Theorem 2 entails a careful analysis of tails of $\tilde{\lambda}_i$ and uses standard subgaussian and subexponential concentration bounds. To bound the error term in (8), we rely on the measurement matrix ensemble satisfying a certain separation condition; we denote this event by E and describe it in detail shortly. Denoting by S the support of λ and by $\hat{S}$ the support of $\hat{\lambda}$, we note that $\Pr\left(\hat{S} \neq S\right)$ can be bounded as

$$\Pr\left(\hat{S} \neq S\right) \leq \Pr\left(\hat{S} \neq S | E\right) + \Pr\left(E^c\right)$$

$$\leq \sum_{i\in S}\Pr\left(\tilde{\lambda}_i < \tau|E\right) + \sum_{i'\in S^c}\Pr\left(\tilde{\lambda}_{i'} \geq \tau|E\right) + \Pr\left(E^c\right). \tag{9}$$

We show that the first two terms in the equation above, involving probabilities conditioned on the event E, can be made small. Also, for the subgaussian measurement ensemble, E occurs with large probability, which in turn implies that the overall error can be made small.

Our approach involves deriving tail bounds for $\tilde{\lambda}_i$ conditioned on the measurement matrices, and then choosing a threshold τ to obtain the desired bound for (9); we derive lower tail bounds for $i \in S$ and upper tail bounds for $i' \in S^c$. . The event E mentioned above corresponds to measurement ensemble being such that we can find this threshold τ that allows us to separate these bounds.

Specifically, note that

$$\tilde{\lambda}_i = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{l\in S} X_{jl}(\Phi_{ji}^\top\Phi_{jl}) + \Phi_{ji}^\top W_j\right)^2,$$

where we used $Y_j = \Phi_j X_j + W_j$. We proceed by observing that conditioned on Φ_1^n, $\tilde{\lambda}_i$ is a sum of independent subexponential random variables. In particular, using basic properties of subexponential random variables and the connection between subgaussian and subexponential random variables (described in Lemmas D.1 and D.2 given in the appendix), we get that conditioned on the measurement matrices Φ_1^n, the random variable $\tilde{\lambda}_i$ is

$$\text{subexp}\left(\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2\right),$$

where c_1 and c_2 are absolute constants and

$$\alpha_{ji}^2 = \begin{cases} \|\Phi_{ji}\|_2^4 + \sum_{l\in S\setminus\{i\}}(\Phi_{jl}^\top\Phi_{ji})^2 + \sigma^2\|\Phi_{ji}\|_2^2, & i\in S,\\ \sum_{l\in S}(\Phi_{jl}^\top\Phi_{ji})^2 + \sigma^2\|\Phi_{ji}\|_2^2, & \text{otherwise.}\end{cases}$$

Using standard tail bounds for subexponential random variables given in Lemma D.1 and denoting $\mu_i \stackrel{\text{def}}{=} \mathbb{E}\left[\tilde{\lambda}_i|\Phi_1^n\right] = \frac{1}{n}\sum_{j=1}^n \alpha_{ji}^2, i\in[d]$, we have for $i \in S$,

$$\Pr\left(\tilde{\lambda}_i < \tau|\Phi_1^n\right) \leq \exp\left(-\min\left\{\frac{n^2(\mu_i-\tau)^2}{c_1\sum_{j=1}^n\alpha_{ji}^4}, \frac{n(\mu_i-\tau)}{c_2\max\limits_{j\in[n]}\alpha_{ji}^2}\right\}\right),$$

and for $i' \in S^c$,

$$\Pr\left(\tilde{\lambda}_{i'} \geq \tau|\Phi_1^n\right) \leq \exp\left(-\min\left\{\frac{n^2(\tau-\mu_{i'})^2}{c_1\sum_{j=1}^n\alpha_{ji'}^4}, \frac{n(\tau-\mu_{i'})}{c_2\max\limits_{j\in[n]}\alpha_{ji'}^2}\right\}\right).$$

We can upper bound the sum of the first two terms in (9) by $\delta/2$ by showing that with large probability Φ_1^n takes values for which we get each term above bounded by roughly $\delta' \stackrel{\text{def}}{=} \delta/(4\max\{(d-k),k\})$. In particular, using a manipulation of the expression for exponents, each of the conditional probabilities above will be less than δ' if τ satisfies the following condition for any $i \in S$ and $i' \in S^c$:

$$\mu_{i'} + \nu_{i'} \leq \tau \leq \mu_i - \nu_i, \tag{10}$$

where

$$\nu_i \stackrel{\text{def}}{=} \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4\log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2\log\frac{1}{\delta'}\right\},$$

and a similar definition holds for $\nu_{i'}$. Thus, the sufficient condition in (10) can be rewritten as

$$\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji}^2 - \frac{1}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2 \log\frac{1}{\delta'}\right\}$$
$$+ \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji'}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji'}^2 \log\frac{1}{\delta'}\right\}. \tag{11}$$

Let E denote the event that for all $i \in S$ and $i' \in S^c$, condition (11) is satisfied by the measurement matrix ensemble. We will show that for Φ_1^n drawn from the subgaussian ensemble satisfying assumption 2, the event E occurs with high probability. We establish this claim by showing that each term in (11) concentrates well around its expected value and roughly $nm^2 \geq ck^2 \log(1/\delta')$ suffices to guarantee that the separation required in (11) holds with large probability. The following result, which we prove in Appendix A, shows that (11) holds with large probability for all pairs $(i, i') \in S \times S^c$.

Lemma 2. *The separation condition* (11) *holds simultaneously for all pairs* $(i, i') \in S \times S^c$ *with probability at least* $1-\delta$ *if* $n \geq c(k/m + \sigma^2)^2 \log(1/\delta')$*, where* $\delta' = \delta/(4\ \max\{k, d-k\})$.

Choosing the probability parameter to be $\delta/2$ in Lemma 2, we see that the third term in (9) can be at most $\delta/2$, leading to an overall error probability of at most δ. Further, noting that $2\log(1/\delta') \geq \log(16k(d-k)/\delta)$, we obtain the result claimed in the theorem. □

Remark 7. As long as the noise variance is sufficiently small, i.e., $\sigma^2 < k/m$, our estimator is sample-optimal and achieves the same scaling as the lower bound that we prove in Section IV.

Remark 8. The separation condition (11) fails to hold for $n = 1$, regardless of which measurement ensemble is used. This is to be expected when $m < k$, since from our lower bound for sample complexity, multiple samples are necessary in the $m < k$ regime.

C. Simulation results

In this section, we numerically evaluate the performance of the closed-form estimator in (5). Our focus will be on exact support recovery and we will study the performance of our estimator over multiple trials. For our experiments we use measurement matrices that are independent across samples and have i.i.d. $\mathcal{N}(0, 1/m)$ entries. To generate measurements, we first pick a support uniformly at random from all possible supports of size k. Next, the data vectors are generated according to one of two methods. In the first method, the nonzero entries of the data have i.i.d. $\mathcal{N}(0,1)$ entries. In the second method, the nonzero entries are i.i.d Rademacher (i.e., $\{+1, -1\}$-valued with equal probability). Both these distributions are subgaussian with variance parameters that are a constant multiple of the respective variances. We generate noiseless measurements Y_1^n according to the linear model described before. For a fixed value of d, k, m and n, we generate multiple instances of the problem and provide it as input to the estimator. For every instance, we declare success or failure depending on whether the support is exactly recovered or not and the success rate is the fraction of instances on which the recovery is successful. For our experiments, we performed 200 trials for every set of parameters. We can see from Figure 1 in Section I that the experimental results closely agree with our predictions. Also, the constant of proportionality is small, roughly between 15 and 20.

We also perform simulations for the case when the measurements are noisy. In particular, we consider noise vectors $W_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I),\ i \in [n]$, and X_1^n Gaussian distributed as described before. We plot the probability of exact support recovery against the normalized number of samples for four different noise values of the noise

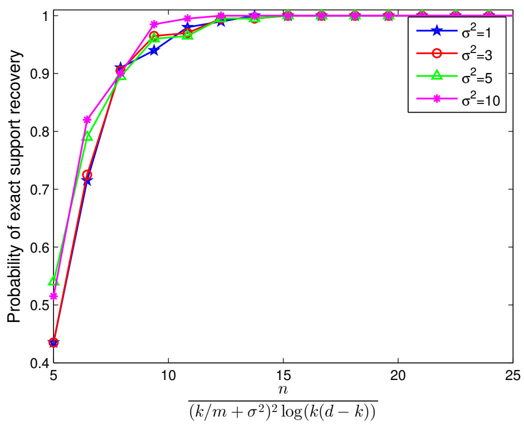


Fig. 2: Performance of the closed-form estimator for different noise levels with $d = 100,\ m = 2,\ k = 10$.

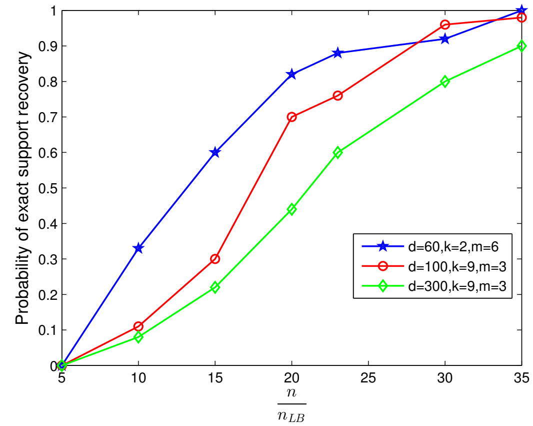


Fig. 3: Performance of MSBL in the noiseless case for different parameter values.

variance, while the other parameters are kept fixed at $d = 100$, $m = 2$, and $k = 10$. It can be seen from Figure 2 that the four curves overlap, indicating that the scaling of n with respect to the noise variance is tight. Finally, Figure 3 shows the performance of MSBL [34], where we plot the probability of exact support recovery against the normalized number of samples (the normalization factor $n_{LB} = (k^2(1-m/k)^4/m^2)\log(k(d-k))$ is from the lower bound established in the next section). It can be seen that the curves do not overlap, indicating that MSBL has a different scaling of n with respect to the parameters m, k, d than what is predicted in the lower bound.

IV. Lower bound

In this section, we prove the lower bound for sample complexity claimed in Theorem 1. We work with the Gaussian setting, where each Φ_i has independent, zero-mean Gaussian entries with variance $1/m$. Denote by S_0 the set $\{1, \ldots, k\}$ and by $S_{i,j}$, $1 \le i \le k < j \le d$, the set obtained by replacing the element i in S_0 with j from S_0^c. Let U be distributed uniformly over the pairs $\{(i, j) : 1 \le i \le k, k+1 \le j \le d\}$. The unknown support is set to be S_U; the random variables X_1^n and linear measurements $Y_i = \Phi_i X_i$ are generated as before.

We consider the Bayesian hypothesis testing problem where we observe Y^n and seek to determine U. Given any support estimator $\hat{S}$, we can use it to find an estimate for the support, which in turn will give an estimate $\hat{U}$ for U. Clearly, $\Pr\left(\hat{U} \ne U\right)$ equals $\Pr\left(\hat{S} \ne S_U\right)$, which must be less than $1/3$ by our assumption. On the other hand, by Fano's inequality, we get

$$\begin{aligned}\Pr\left(\hat{U} \ne U\right) &\ge 1 - \frac{I(Y_1^n; U) + 1}{\log(k(d-k))} \\ &\ge 1 - \frac{\max_u D(\mathrm{P}_{Y^n|S_u} \| \mathrm{P}_{Y^n|S_0}) + 1}{\log(k(d-k))},\end{aligned}$$

where $\mathrm{P}_{Y^n|S}$ denotes the distribution of the measurements when the support of λ is S (a proof for the second inequality can be found in [10, Theorem 21]). Note that $\mathrm{P}_{Y^n|S} = \prod_{i=1}^n \mathrm{P}_{Y_i|S}$ with each $\mathrm{P}_{Y_i|S}$ having the same distribution which we denote by $\mathrm{P}_{Y|S}$. Thus, $D(\mathrm{P}_{Y^n|S_u} \| \mathrm{P}_{Y^n|S_0}) = nD(\mathrm{P}_{Y|S_u} \| \mathrm{P}_{Y|S_0})$.

Next, we bound $D(\mathrm{P}_{Y|S_u} \| \mathrm{P}_{Y|S_0})$. Denote by Φ_S the $m \times k$ submatrix of Φ obtained by restricting to the columns in S and by A_S the Gram matrix $\Phi_S \Phi_S^\top$ of Φ_S. Further, let $a_1 \ge \ldots \ge a_m > 0$ and $b_1 \ge \ldots \ge b_m > 0$ be the respective eigenvalues of A_{S_u} and A_{S_0}. Note that $a_m > 0$ and $b_m > 0$ hold with probability 1 since $m \le k$.

Denoting by $\mathrm{P}_{Y|S,\Phi}$ the conditional distribution of the measurement when the measurement matrix is fixed to Φ, we get

$$\begin{aligned} D\big(\mathrm{P}_{Y|S_u,\Phi}\|\mathrm{P}_{Y|S_0,\Phi}\big) &= \frac{1}{2}\left(\log\frac{|A_{S_0}|}{|A_{S_u}|} + \mathrm{Tr}(A_{S_0}^{-1}A_{S_u}) - m\right) \\ &\leq \frac{1}{2}\sum_{i=1}^{m}\left(\log\frac{b_i}{a_i} - \left(1-\frac{a_i}{b_i}\right)\right) \\ &\leq \frac{1}{2}\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_ib_i}, \end{aligned}$$

where in the first inequality holds by Lemma D.5 and the second inequality holds since $\log x + (1-x)/x \leq (x-1)^2/x$ for all $x > 0$. Using convexity of the KL divergence, we can get

$$\begin{aligned} D\big(\mathrm{P}_{Y|S_u}\|\mathrm{P}_{Y|S_0}\big) &\leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_ib_i}\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_mb_m}\right] \end{aligned}$$

Note that the expression on the right does not depend on our choice of u; we fix $u = (1, k+1)$. With an abuse of notation, we denote by Φ_j the jth column of a random matrix Φ with independent $\mathcal{N}(0, 1/m)$ distributed entries. Using the Cauchy-Schwarz inequality twice, we get

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_m^2}\right] &\leq \sqrt{\mathbb{E}\left[\frac{1}{a_m^2b_m^2}\right]}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{m}(a_i-b_i)^2\right)^2\right]} \\ &\leq \sqrt{\mathbb{E}\left[\frac{1}{a_m^4}\right]}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{m}(a_i-b_i)^2\right)^2\right]}, \end{aligned}$$

where in the second inequality we also used the fact that a_is and b_is are identically distributed. The Hoffman-Wielandt inequality[3] [15] can be used to handle the second term on the right-side. In particular, we have $\sum_{i=1}^{m}(a_i - b_i)^2 \leq \|A_{S_0} - A_{S_u}\|_F^2$ where the right-side coincides with $\|\Phi_1\Phi_1^\top - \Phi_{k+1}\Phi_{k+1}^\top\|_F^2$ since $u = (1, k+1)$. Using the triangle inequality for Frobenius norm and noting that $\|\Phi_i\Phi_i^\top\|_F$ equals $\|\Phi_i\|_2^2$ for a vector Φ_i, we get

$$\mathbb{E}\left[\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_m^2}\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{a_m^4}\right]}\sqrt{\mathbb{E}\left[(\|\Phi_1\|_2^2 + \|\Phi_{k+1}\|_2^2)^4\right]}.$$

Recall that Φ_1 and Φ_{k+1} are independent $\mathcal{N}(0, \frac{1}{m}I)$ distributed random vectors, and therefore $m(\|\Phi_1\|_2^2 + \|\Phi_{k+1}\|_2^2)$ is a chi-squared random variable with $2m$ degrees of freedom. Using the expression for the fourth moment of a chi-squared random variable gives us

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_m^2}\right] &\leq \sqrt{\mathbb{E}\left[\frac{1}{a_m^4}\right]}\sqrt{\frac{1}{m^4}\frac{(m+3)!}{(m-1)!}} \\ &\leq c'\sqrt{\mathbb{E}\left[\frac{1}{a_m^4}\right]} \end{aligned}$$

where c' is an absolute constant.

It only remains to bound $\mathbb{E}\left[1/a_m^4\right]$, where a_m is the minimum eigenvalue of the $(m \times m)$ Wishart matrix A_{S_u}.

[3] For normal matrices A and B with spectra $\{a_i\}$ and $\{b_i\}$, there exists a permutation π of $[n]$ such that $\sum_i (a_{\pi(i)} - b_i)^2 \leq \|A - B\|_F^2$. When A and B are p.s.d, the left-side is minimum when both sets of eigenvalues are arranged in increasing (or decreasing) order.

Using Lemma C.1 in Appendix C, we can obtain

$$\mathbb{E}\left[a_m^{-4}\right] \leq \frac{c'' m^4}{k^4(1-m/k)^8}.$$

By combining all the steps above, we get

$$\frac{1}{3} \geq \Pr\left(\hat{S} \neq S_U\right) \geq 1 - \frac{\frac{cnm^2}{k^2(1-m/k)^4} + 1}{\log k(d-k)},$$

for a constant c. Observing that the $(1-m/k)^4$ term can be absorbed into c when $m < k/2$ yields the desired bound.

V. Extension to nonbinary variances

In this section, we extend our results to the case where λ is not necessarily binary. Specifically, we have the following assumption.

Assumption 3. *The entries of X_i, $i \in [n]$, are independent and zero-mean with $\mathbb{E}\left[X_{i,j}^2\right] = \lambda_j$ for $\lambda \in \{u \in \mathbb{R}^d : \|u\|_0 = k, \lambda_{min} \leq u_i \leq \lambda_{max}\}$ and $X_{i,j} \sim subG(c\lambda_j)$, where $0 < \lambda_{min} \leq \lambda_{max}$ and c is an absolute constant. In addition, we assume that $\lambda_{min}/\lambda_{max} > k/(k+m-1)$.*

Our sample complexity result continues to hold with an additional scaling by a factor of $\lambda_{\text{max}}^2/\lambda_{\text{min}}^2$. In particular, we have the following result, where we limit to the noiseless setting.

Theorem 3. *For $\sigma^2 = 0$, the sample complexity of support recovery under Assumption 2 and Assumption 3 satisfies*

$$c\frac{\lambda_{max}^2}{\lambda_{min}^2}\frac{k^2}{m^2}\log(d-k+1) \leq n^*(m,k,d) \leq C\frac{\lambda_{max}^2}{\lambda_{min}^2}\left(\frac{k}{m}+1\right)^2 \log\left(\frac{k(d-k)}{\delta}\right),$$

provided $m \geq (\log k)^2$, with c and C being absolute constants.

The techniques used for proving the upper and lower bounds remain essentially the same, and we highlight the key changes in the next two subsections.

We start by extending the bias calculation in Lemma 1 to the more general nonbinary setting.

Lemma 3. *Let the estimator $\tilde{\lambda}$ be as defined in (4). Then, under Assumptions 1 and 2 with $c = c' = c'' = 1$, we have that*

$$\mathbb{E}\left[\tilde{\lambda}_i\right] = \frac{m+1}{m}\lambda_i + \frac{1}{m}Tr(K_\lambda) + \sigma^2, \quad i \in [d],$$

where the expectation is with respect to the joint distribution of (X_1^n, Φ_1^n, W_1^n).

Proof. We recall from the proof of Lemma 1 that our estimator can be rewritten in the following form:

$$\tilde{\lambda}_i = \frac{1}{n}\sum_{j=1}^n \left(\sum_{l\in S} X_{jl}(\Phi_{ji}^\top \Phi_{jl}) + \Phi_{ji}^\top W_j\right)^2, \quad i \in [d].$$

Noting that $\mathbb{E}\left[X_{jl}^2\right] = \lambda_l$ for all $l \in S$, $j \in [n]$ and that $\mathbb{E}\left[X_{jl}^2\right] = 0$ for all $l \in S^c$, $j \in [n]$, we have

$$\mathbb{E}\left[\tilde{\lambda}_i|\Phi_1^n\right] = \begin{cases} \frac{1}{n}\sum_{j=1}^n \left(\lambda_i\|\Phi_{ji}\|_2^4 + \sum_{l\in S\setminus\{i\}}\lambda_l(\Phi_{ji}^\top\Phi_{jl})^2 + \sigma^2\|\Phi_{ji}\|_2^2\right), & \text{if } i \in S, \\ \frac{1}{n}\sum_{j=1}^n \left(\sum_{l\in S}\lambda_l(\Phi_{ji}^\top\Phi_{jl})^2 + \sigma^2\|\Phi_{ji}\|_2^2\right), & \text{otherwise.} \end{cases}$$

Taking expectation with respect to Φ_1^n and using Lemma D.4 gives

$$\mathbb{E}\left[\tilde{\lambda}_i\right] = \frac{m+1}{m}\lambda_i + \frac{1}{m}\text{Tr}(K_\lambda) + \sigma^2, \quad i \in [d].$$

□

Remark 9. Note that when λ is binary, $\text{Tr}(K_\lambda) = k$ and the result above reduces to Lemma 1.

Our final estimate for the support is the same as before, namely, it computes $\tilde{\lambda}$ and declares the indices of the k largest entries as the support. However, as before, we work with a threshold based estimator, with the bias terms in Lemma 3 being accounted for in the threshold.

Following the same series of arguments as in the binary case, we can show that to achieve a small probabilty of error, it suffices if a separation condition similar to (11) holds. Focusing on the $\sigma^2 = 0$ case and using the assumption that $\lambda_i \in [\lambda_{\text{min}}, \lambda_{\text{max}}]$, we have that for $\Pr\left(\tilde{S} \neq S\right) \leq \delta$, it suffices if the following condition holds for every $i \in S$ and every $i' \in S^c$:

$$\frac{\lambda_{\text{min}}}{n}\sum_{j=1}^{n}\alpha_{ji}^2 - \frac{\lambda_{\text{max}}}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq \lambda_{\text{max}}\Bigg(\max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2\log\frac{1}{\delta'}\right\} + \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji'}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji'}^2\log\frac{1}{\delta'}\right\}\Bigg),$$

where $\delta' = \delta/(2\max\{k, d-k\})$. Incorporating the scaling due to λ_{min} and λ_{max} into our concentration bounds in the proof of Lemma 2, and simplifying, we get that

$$n \geq C\,\frac{\lambda_{\text{max}}^2}{\lambda_{\text{min}}^2}\left(\frac{k}{m}+1\right)^2 \log\left(\frac{k(d-k)}{\delta}\right)$$

samples suffice for $\Pr\left(\tilde{S} \neq S\right) \leq \delta$, provided $\lambda_{\text{min}}/\lambda_{\text{max}} > k/(k+m-1)$.

Remark 10. One can see from the result above that the number of samples increases as λ_{min} decreases. This is because if the variance along any coordinate $i \in S$ is very small, then it is difficult for the estimator to distinguish it from zero. On the other hand, a large λ_{max} can cause "faulty inclusion" of coordinates in the support set. This is because if we are estimating the variance along a coordinate i that is not in the support and if the variance along some coordinate j is large, then even a small amount of correlation between columns i and j of the measurement matrices can result in coordinate i being classified as belonging to the support.

For the case when the measurements are noisy, a similar calculation shows that

$$n \geq c\,\frac{\lambda_{\text{max}}^2}{\lambda_{\text{min}}^2}\left(\frac{k}{m}+1+\frac{\sigma^2}{\lambda_{\text{max}}}\right)^2 \log\left(\frac{k(d-k)}{\delta}\right)$$

samples are sufficient.

A. *Lower bound*

We assume that the unknown λ is uniformly distributed over the set $\{\lambda_0, \lambda_1, \ldots, \lambda_{d-k}\}$, with $\lambda_i \in \mathbb{R}^d$. The jth entry of λ_i, denoted λ_{ij}, is given by

$$\lambda_{ij} = \begin{cases} \lambda_{\text{max}}, & \text{if } j \in [k-1], \\ \lambda_{\text{min}}, & \text{if } j = k+i, \\ 0, & \text{otherwise.} \end{cases}$$

for any $i \in \{0, 1, \ldots, d-k\}$.

Our goal is to characterize the KL divergence between distributions on the measurements arising from two different λs in the set we described above, one of which we fix as λ_0. Computing this divergence as before, we see that

$$D\big(\mathrm{P}_{Y|\lambda} \| \mathrm{P}_{Y|\lambda_0}\big) \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \frac{(a_i' - b_i')^2}{(a_m')^2}\right], \tag{12}$$

where $\{a_i'\}_{i=1}^m$ and $\{b_i'\}_{i=1}^m$ denote the eigenvalues of $A_\lambda \stackrel{\text{def}}{=} \Phi K_\lambda \Phi^\top$ and $A_{\lambda_0} \stackrel{\text{def}}{=} \Phi K_{\lambda_0} \Phi^\top$ respectively. Noting that $\sum_{i=1}^m (a_i' - b_i')^2 \leq \|A_\lambda - A_{\lambda_0}\|_F^2 = \lambda_{\text{min}}^2 \|\Phi_1\Phi_1^\top - \Phi_{k+1}\Phi_{k+1}^\top\|_F^2$, an application of the Hoffman-Wielandt inequality yields

$$\mathbb{E}\left[\sum_{i=1}^{m} \frac{(a_i' - b_i')^2}{(a_m')^2}\right] \leq c\lambda_{\text{min}}^2 \sqrt{\mathbb{E}\left[\frac{1}{(a_m')^4}\right]}. \tag{13}$$

Recall that from Lemma C.1, we have a bound on the fourth moment of the smallest eigenvalue a_m of $A_S = \Phi_S\Phi_S$ for $S \subseteq [d]$. We now try to relate a_m' and a_m. We start by noting that

$$\begin{aligned} A_\lambda &= \lambda_{\text{max}} \sum_{i=1}^{k-1} \Phi_i\Phi_i^\top + \lambda_{\text{min}}\Phi_{k+1}\Phi_{k+1}^\top \\ &\succcurlyeq \lambda_{\text{max}} \sum_{i=1}^{k-1} \Phi_i\Phi_i^\top, \end{aligned}$$

where $A \succcurlyeq B$ if $A - B$ is a positive semi-definite matrix. The above inequality in turn gives a relation between the eigenvalues of A_λ and those of $\lambda_{\text{max}} \sum_{i=1}^{k-1} \Phi_i\Phi_i^\top$. In particular, for the minimum eigenvalue, we have $a_m' \geq \lambda_{\text{max}} a_m$. Combining this fact with the inequalities in (12) and (13), and using Lemma C.1, we get

$$\begin{aligned} D\big(\mathrm{P}_{Y|\lambda} \| \mathrm{P}_{Y|\lambda_0}\big) &\leq c' \frac{\lambda_{\text{min}}^2}{\lambda_{\text{max}}^2} \sqrt{\mathbb{E}\left[\frac{1}{a_m^4}\right]} \\ &\leq c'' \frac{\lambda_{\text{min}}^2}{\lambda_{\text{max}}^2} \frac{m^2}{(k-1)^2} \left(1 - \frac{m}{k}\right)^{-4}. \end{aligned}$$

This is roughly the same bound as in the binary case, except with an additional scaling by a factor of $\lambda_{\text{min}}^2/\lambda_{\text{max}}^2$. As a consequence of this, we can show, using similar calculations as before, that if

$$n \leq c \frac{\lambda_{\text{max}}^2}{\lambda_{\text{min}}^2} \frac{k^2}{m^2} \left(1 - \frac{m}{k}\right)^4 \log(d - k + 1),$$

then the error probability $\Pr\left(\hat{S} \neq \mathtt{supp}(\lambda_0)\right) \geq 1/3$.

VI. Concluding Remarks

Our sample complexity result implies that independent measurements applied to the same sample are much more helpful than those applied to independent samples. There are several possible extensions of our results.

We have shown that covariance-based methods provide a reliable way to recover the support in the measurement-constrained setting of $m < k$, where traditional sparse recovery methods do not work. This framework can accommodate other kinds of structures on the covariance matrix; one can simply project the closed form estimate onto an appropriate constraint set and then use this estimate to make inferences about the data. It would be interesting to explore this in more detail.

One can consider using the same measurement matrix for all samples. In this case, we observe empirically that our estimate does not perform well, but we do not have a complete theoretical understanding of this phenomenon for our setting (see [3, Proposition 2] for a related discussion). Our current results are tight only for the high SNR case of $\sigma^2 < k/m$; it will be of interest to derive lower bounds for the noisy setting.

For the case where the samples do not share a common support but instead have supports drawn from a small set of allowed supports, our current estimator can be used to recover the union of the allowed supports. Designing estimators for segregating the union into individual supports and labeling the samples based on the supports is an interesting direction for further work.

Appendix A
Proof of Lemma 2

We recall the statement of Lemma 2 here for easy reference.

Lemma A.1. *For all pairs $(i, i') \in S \times S^c$, the separation condition*

$$\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji}^2 - \frac{1}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2 \log\frac{1}{\delta'}\right\}$$
$$+ \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji'}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji'}^2 \log\frac{1}{\delta'}\right\} \tag{14}$$

holds with probability at least $1-\delta$ if $n \geq c(k/m+\sigma^2)^2\log(1/\delta')$ and $m \geq (\log k)^2$, where $\delta' = \delta/(4\max\{k, d-k\})$.

Proof. The proof involves studying the tail behaviour of each term in (14). In particular, we derive a lower bound on the first term and upper bounds on the remaining terms that hold with high probability over the subgaussian measurement ensemble, and establish conditions under which the separation in (14) holds for a fixed pair (i, i'). A union bound over all $k(d-k)$ pairs then gives us the result claimed in the lemma. For clarity of presentation, details of the tail bounds for each term in (14) are presented in Appendix B, which in turn build on standard concentration bounds for subgaussian and subexponential random variables reviewed as preliminaries in Appendix D. Also, while analyzing each term in (14), we use the same symbol μ to denote the expectation of that term to keep notation simple. Similarly, the definitions of terms like μ_1, μ_2 and μ_3 will be clear from the context.

For the first term on the left side of (14), we study the behaviour of its left tail. That is, we look at $\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji}^2 \leq \mu - t\right)$, where recall

$$\alpha_{ji}^2 = \|\Phi_{ji}\|_2^4 + \sum_{l\in S\setminus\{i\}}(\Phi_{jl}^\top\Phi_{ji})^2 + \sigma^2\|\Phi_{ji}\|_2^2, \tag{15}$$

and

$$\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji}^2\right].$$

We will denote the expected values of the three terms in (15) by μ_1, μ_2 and μ_3, respectively. Note that

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji}^2 \leq \mu - t\right) \leq \Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|\Phi_{ji}\|_2^4 \leq \mu_1 - t_1\right)$$
$$+ \Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{t\in S\setminus\{i\}}(\Phi_{jt}^\top\Phi_{ji})^2 \leq \mu_2 - t_2\right)$$

$$+ \Pr\left(\frac{\sigma^2}{n}\sum_{j=1}^{n}\|\Phi_{ji}\|_2^2 \le \mu_3 - t_3\right), \tag{16}$$

where $t = t_1 + t_2 + t_3$ and $\mu = \mu_1 + \mu_2 + \mu_3$. Since Φ_{ji} has independent subgaussian entries with variance $1/m$, we can see from Lemmas D.2 and D.3 that $\|\Phi_{ji}\|_2^2$ is subexponential with parameters $(c_1/m, c_2/m)$ and that $(\sigma^2/n)\sum_{j=1}^{n}\|\Phi_{ji}\|_2^2$ is subexponential with parameters $(c_1\sigma^4/mn, c_2\sigma^2/mn)$. Thus, the third term in (16) can be handled using subexponential concentration bound from Lemma B.1, which gives

$$\Pr\left(\frac{\sigma^2}{n}\sum_{j=1}^{n}\|\Phi_{ji}\|_2^2 \le \mu_3 - t_3\right) \le \frac{\varepsilon}{3}$$

for

$$\mu_3 - t_3 = \sigma^2\left(1 - \max\left\{\sqrt{\frac{c_1}{mn}\log\frac{3}{\varepsilon}}, \frac{c_2}{mn}\log\frac{3}{\varepsilon}\right\}\right).$$

For the first and second terms on the right side of (16), we provide details in Lemmas B.2 and B.4. In particular,

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|\Phi_{ji}\|_2^4 \le \mu_1 - t_1\right) \le \frac{\varepsilon}{3}$$

for

$$\mu_1 - t_1 = \min\left\{\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{3}{\varepsilon}}\right)^2, \left(1 - \frac{c_2}{mn}\log\frac{3}{\varepsilon}\right)^2\right\},$$

and when $n \ge (c_2^2/c_1)\log(12/\varepsilon)$,

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{t\in S\setminus\{i\}}(\Phi_{jt}^\top\Phi_{ji})^2 \le \mu_2 - t_2\right) \le \frac{\varepsilon}{3}$$

for

$$\begin{aligned}\mu_2 - t_2 =& \frac{k-1}{m}\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{12}{\varepsilon}}\right) \\ &- \sqrt{\frac{1}{mn}\log\frac{12}{\varepsilon}}\max\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\max\left\{\left(1 + \sqrt{\frac{c_1}{m}\log\frac{12n}{\varepsilon}}\right), \left(1 + \frac{c_1}{m}\log\frac{12n}{\varepsilon}\right)\right\}.\end{aligned}$$

Using these results in (16), we see that

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji}^2 \le \mu - t\right) \le \varepsilon,$$

for

$$\begin{aligned}\mu - t = &\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{3}{\varepsilon}}\right)^2 + \frac{k-1}{m}\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{12}{\varepsilon}}\right) \\ &- \sqrt{\frac{1}{mn}\log\frac{12}{\varepsilon}}\max\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\max\left\{\left(1 + \sqrt{\frac{c_1}{m}\log\frac{12n}{\varepsilon}}\right), \left(1 + \frac{c_2}{m}\log\frac{12n}{\varepsilon}\right)\right\} \\ &+ \sigma^2\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{3}{\varepsilon}}\right),\end{aligned} \tag{17}$$

when $n \ge (c_2^2/c_1)\log(12/\varepsilon)$.

We now consider the second term on the left side of (14), and study the behaviour of its right tail. That is, we look at $\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq \mu + t\right)$ for $i' \in S^c$, where

$$\alpha_{ji'}^2 = \sum_{l\in S}(\Phi_{jl}^\top \Phi_{ji'})^2 + \sigma^2 \|\Phi_{ji'}\|_2^2.$$

Letting $\mu_1 = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{l\in S}(\Phi_{jl}^\top \Phi_{ji'})^2\right]$ and $\mu_2 = \mathbb{E}\left[\frac{\sigma^2}{n}\sum_{j=1}^{n}\|\Phi_{ji'}\|_2^2\right]$, we note that

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq \mu + t\right) \leq \Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l\in S}(\Phi_{jl}^\top \Phi_{ji'})^2 \geq \mu_1 + t_1\right) + \Pr\left(\frac{\sigma^2}{n}\sum_{j=1}^{n}\|\Phi_{ji'}\|_2^2 \geq \mu_2 + t_2\right),$$

where $t = t_1 + t_2$. As before, we use Lemma B.1 to get

$$\Pr\left(\frac{\sigma^2}{n}\sum_{j=1}^{n}\|\Phi_{ji'}\|_2^2 \geq \mu_2 + t_2\right) \leq \frac{\varepsilon}{2}$$

for

$$\mu_2 + t_2 = \sigma^2\left(1 + \max\left\{\sqrt{\frac{c_1}{mn}\log\frac{2}{\varepsilon}}, \frac{c_2}{mn}\log\frac{2}{\varepsilon}\right\}\right),$$

and Lemma B.5 to get

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l\in S}(\Phi_{jl}^\top \Phi_{ji'})^2 \geq \mu_1 + t_1\right) \leq \frac{\varepsilon}{2}$$

for

$$\begin{aligned}\mu_1 + t_1 =& \frac{k}{m}\left(1 + \sqrt{\frac{c_1}{mn}\log\frac{8}{\varepsilon}}\right) \\ &+ \sqrt{\frac{1}{mn}\log\frac{8}{\varepsilon}}\ \max\left\{\sqrt{c_1\frac{k}{m}}, c_2\right\}\max\left\{\left(1 + \sqrt{\frac{c_1}{m}\log\frac{8n}{\varepsilon}}\right), \left(1 + \frac{c_2}{m}\log\frac{8n}{\varepsilon}\right)\right\},\end{aligned}$$

when $n \geq (c_2^2/c_1)\log(8/\varepsilon)$. Putting these results together, we get

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq \mu + t\right) \leq \varepsilon,$$

for

$$\begin{aligned}\mu + t =& \frac{k}{m}\left(1 + \sqrt{\frac{c_1}{mn}\log\frac{8}{\varepsilon}}\right) \\ &+ \sqrt{\frac{1}{mn}\log\frac{8}{\varepsilon}}\ \max\left\{\sqrt{c_1\frac{k}{m}}, c_2\right\}\max\left\{\left(1 + \sqrt{\frac{c_1}{m}\log\frac{8n}{\varepsilon}}\right), \left(1 + \frac{c_2}{m}\log\frac{8n}{\varepsilon}\right)\right\} \\ &+ \sigma^2\left(1 + \sqrt{\frac{c_1}{mn}\log\frac{2}{\varepsilon}}\right),\end{aligned} \tag{18}$$

when $n \geq (c_2^2/c_1)\log(8/\varepsilon)$.

For the third term in (14), namely, $\max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4\log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2\log\frac{1}{\delta'}\right\}$, we consider the possibility of either argument attaining the maximum and study the respective right tails. First, we look at $\Pr\left(\sqrt{\frac{1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4} \geq \mu + t\right)$

for $i \in S$. We note that by the union bound,

$$
\begin{aligned}
\Pr\left(\sqrt{\frac{1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4} \geq \mu+t\right) &= \Pr\left(\sum_{j=1}^{n}\alpha_{ji}^4 \geq n^2(\mu+t)^2\right) \\
&\leq \sum_{j=1}^{n}\Pr\left(\alpha_{ji}^4 \geq n(\mu+t)^2\right) \\
&= \sum_{j=1}^{n}\Pr\left(\alpha_{ji}^2 \geq \sqrt{n}(\mu+t)\right) \\
&\leq n\Pr\left(\|\Phi_{1i}\|_2^4 \geq \frac{\sqrt{n}}{3}(\mu+t)\right) + n\Pr\left(\sum_{l\in S\setminus\{i\}}(\Phi_{1i}^\top\Phi_{1l})^2 \geq \frac{\sqrt{n}}{3}(\mu+t)\right) \\
&\quad + n\Pr\left(\sigma^2\|\Phi_{1i}\|_2^2 \geq \frac{\sqrt{n}}{3}(\mu+t)\right). \qquad (19)
\end{aligned}
$$

We use Lemma D.1 for the first and third terms and Lemma B.5 for the second term. A direct application of Lemma B.5 with $n=1$ for the second term however requires the assumption that $m \geq (c_2^2/c_1)\log(12n/\varepsilon)$ (note that the second term in (19) needs to be upper bounded by $\varepsilon/3n$). While in our setting such an assumption on n is acceptable, we would like to avoid making this assumption on m at this stage. We therefore omit the simplification done at the end of Lemma B.5 to get

$$
\Pr\left(\sqrt{\frac{1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4} \geq \mu+t\right) \leq \varepsilon
$$

for

$$
\begin{aligned}
\mu+t = &\frac{3}{\sqrt{n}}\left(1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{3n}{\varepsilon}}, \frac{c_2}{m}\log\frac{3n}{\varepsilon}\right\}\right)^2 \\
&+\frac{3}{\sqrt{n}}\left(\frac{k-1}{m}+\max\left\{\frac{c_2}{m}\log\frac{9n}{\varepsilon}, \sqrt{c_1\frac{k-1}{m^2}\log\frac{9n}{\varepsilon}}\right\}\right)\left(1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{9n}{\varepsilon}}, \frac{c_2}{m}\log\frac{9n}{\varepsilon}\right\}\right) \\
&+\frac{3\sigma^2}{\sqrt{n}}\left(1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{3n}{\varepsilon}}, \frac{c_2}{m}\log\frac{3n}{\varepsilon}\right\}\right). \qquad (20)
\end{aligned}
$$

Next, we look at $\Pr\left(\max_{j\in[n]}\alpha_{ji}^2 \geq \mu+t\right)$ for $i \in S$. We notice that by the union bound, we have

$$
\begin{aligned}
\Pr\left(\max_{j\in[n]}\alpha_{ji}^2 \geq \mu+t\right) &\leq \sum_{j=1}^{n}\Pr\left(\alpha_{ji}^2 \geq \mu+t\right) \\
&\leq \sum_{j=1}^{n}\Bigg[\Pr\left(\|\Phi_{ji}\|_2^4 \geq \frac{\mu+t}{3}\right) + \Pr\left(\sum_{l\in S\setminus\{i\}}(\Phi_{ji}^\top\Phi_{jl})^2 \geq \frac{\mu+t}{3}\right) \\
&\qquad + \Pr\left(\sigma^2\|\Phi_{ji}\|_2^2 \geq \frac{\mu+t}{3}\right)\Bigg]. \qquad (21)
\end{aligned}
$$

We now handle each of the three terms on the right-side of (21) separately. We will use Lemma B.1 for the first and third terms and Lemma B.4 for the second term. In particular, for every $j \in [n]$, we have that

$$
\Pr\left(\|\Phi_{ji}\|_2^4 \geq \frac{\mu+t}{3}\right) \leq \varepsilon,
$$

for

$$\mu + t = 3\left(1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{1}{\varepsilon}}, \frac{c_2}{m}\log\frac{1}{\varepsilon}\right\}\right)^2,$$

and that

$$\Pr\left(\|\Phi_{1i}\|_2^2 \geq \frac{\mu+t}{3\sigma^2}\right) \leq \varepsilon$$

for

$$\mu + t = 3\sigma^2\left(1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{1}{\varepsilon}}, \frac{c_2}{m}\log\frac{1}{\varepsilon}\right\}\right).$$

For the second term, we have that for every $j \in [n]$,

$$\Pr\left(\sum_{l \in S\setminus\{i\}} (\Phi_{ji}^\top \Phi_{jl})^2 \geq \frac{\mu+t}{3}\right) \leq \varepsilon,$$

for

$$\mu + t = 3\left(\frac{k-1}{m} + \sqrt{c_2\frac{k-1}{m^2}\log\frac{3}{\varepsilon}}\right)\left(1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{3}{\varepsilon}}, \frac{c_2}{m}\log\frac{3}{\varepsilon}\right\}\right).$$

Substituting these bounds into (21), we get

$$\Pr\left(\max_{j\in[n]} \alpha_{ji}^2 \geq \mu + t\right) \leq 3n\varepsilon,$$

for

$$\mu + t = 3(1 + f(m,\varepsilon)) \cdot \max\left\{\sigma^2, 1 + f(m,\varepsilon), \frac{k-1}{m} + \sqrt{c_2\frac{k-1}{m^2}\log\frac{3}{\varepsilon}}\right\}$$

where $f(m,\varepsilon) = \max\{\sqrt{\frac{c_1}{m}\log\frac{3}{\varepsilon}}, \frac{c_2}{m}\log\frac{3}{\varepsilon}\}$. That is,

$$\Pr\left(\frac{1}{n}\max_{j\in[n]} \alpha_{ji}^2 \geq \mu + t\right) \leq \varepsilon$$

for

$$\mu + t = \frac{3}{n}(1 + f(m,\varepsilon/3n)) \cdot \max\left\{\sigma^2, 1 + f(m,\varepsilon/3n), \frac{k-1}{m} + \sqrt{c_2\frac{k-1}{m^2}\log\frac{9n}{\varepsilon}}\right\}. \tag{22}$$

Comparing (20) and (22), we see that $\frac{1}{n}\max_{j\in[n]} \alpha_{ji}^2$ is $O(k/mn + \sigma^2/n)$ which decays faster with respect to n compared to $\sqrt{\frac{1}{n^2}\sum_{j=1}^{n} \alpha_{ji}^4}$, which is $O(k/m\sqrt{n} + \sigma^2/\sqrt{n})$. Thus, the third term in (14) is dominated by the $O(k/m\sqrt{n} + \sigma^2/\sqrt{n})$ term, which is what we retain in our subsequent calculations.

Finally, for the fourth term in (14), we first look at $\Pr\left(\sqrt{(1/n^2)\sum_{j=1}^{n} \alpha_{ji'}^4} \geq \mu + t\right)$ for $i' \in S^c$. Using similar arguments as in the previous calculation, we get

$$\Pr\left(\sqrt{\frac{1}{n^2}\sum_{j=1}^{n} \alpha_{ji'}^4} \geq \mu + t\right) \leq \varepsilon,$$

for

$$\mu + t = \frac{2}{\sqrt{n}}\left(\frac{k}{m} + \max\left\{\frac{c_2}{m}\log\frac{6n}{\varepsilon}, \sqrt{c_1\frac{k}{m^2}\log\frac{6n}{\varepsilon}}\right\}\right)\left(1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{6n}{\varepsilon}}, \frac{c_2}{m}\log\frac{6n}{\varepsilon}\right\}\right)$$

$$+\frac{2\sigma^2}{\sqrt{n}}\left(1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{2n}{\varepsilon}},\frac{c_2}{m}\log\frac{2n}{\varepsilon}\right\}\right). \tag{23}$$

The $\frac{1}{n}\max_{j\in[n]}\alpha_{ji'}^2$ term, as we discussed before, will lead to a $O(k/mn)$ factor, which can be ignored.

The foregoing calculations provide bounds on each of the four terms occuring in (14), that hold with high probability. We note from (17) and (18) that the left-side of (14) is lower bounded by

$$1-\frac{1}{m}-2\sqrt{\frac{c_1}{mn}\log\frac{24}{\varepsilon}}\left(1+\sigma^2+\frac{k}{m}\right)+\frac{c_1}{mn}\log\frac{6}{\varepsilon}$$
$$-2\sqrt{\frac{c_1k}{m^2n}\log\frac{24}{\varepsilon}}\left(1+\frac{c_2}{m}\log\frac{24n}{\varepsilon}\right) \tag{24}$$

with probability at least $1-\varepsilon$ and from (20) and (23) that the right-side is upper bounded by

$$5\sqrt{\frac{c_1}{n}\log\frac{1}{\delta'}}\Bigg[\left(1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{6n}{\varepsilon}},\frac{c_2}{m}\log\frac{6n}{\varepsilon}\right\}\right)^2$$
$$+\left(\frac{k}{m}+\max\left\{\frac{c_2}{m}\log\frac{18n}{\varepsilon},\sqrt{c_1\frac{k}{m^2}\log\frac{18n}{\varepsilon}}\right\}\right)\left(1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{18n}{\varepsilon}},\frac{c_2}{m}\log\frac{18n}{\varepsilon}\right\}\right)$$
$$+\sigma^2\left(1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{6n}{\varepsilon}},\frac{c_2}{m}\log\frac{6n}{\varepsilon}\right\}\right)\Bigg], \tag{25}$$

with probability at least $1-\varepsilon$. To ensure that (14) holds with probability at least $1-\varepsilon$ for a fixed $(i,i')\in S\times S^c$, we need that (24) exceeds (25). For further simplification, we assume m to be sufficiently large to handle the $\log n$ terms. This assumption on m can possibly be removed by handling the sum in Lemma B.3 and (19) directly and not using the union bound. Choosing $\varepsilon=\delta/(4\,k(d-k))$ to account for the union bound over all (i,i') pairs and focusing on the $n=O((k/m+1+\sigma^2)^2\log(1/\delta'))$ regime, we see that (24) exceeds (25) and separation holds if [4] $m\geq(\log k)^2$.

Thus,

$$n\geq c\left(\frac{k}{m}+1+\sigma^2\right)^2\log\frac{1}{\delta'}$$

samples suffice to ensure separation between the typical values and to guarantee that (14) holds with probability at least $1-\delta$.

□

APPENDIX B
KEY TECHNICAL LEMMAS

Lemma B.1. *Let $Z_1,\ldots,Z_n$ be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Then, there exist absolute constants c_1 and c_2 such that for any $t>0$,*

$$\Pr\left(\left|\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2-1\right|\geq t\right)\leq 2\exp\left(-\mathit{min}\left\{\frac{mn}{c_1}t^2,\frac{mn}{c_2}t\right\}\right).$$

Equivalently, for any $\epsilon>0$,

$$\Pr\left(\left|\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2-1\right|\geq \mathit{max}\left\{\sqrt{\frac{c_1}{mn}\log\frac{2}{\varepsilon}},\frac{c_2}{mn}\log\frac{2}{\varepsilon}\right\}\right)\leq\varepsilon.$$

[4]We use this condition to show that $(1/\sqrt{m})\log(k/m)\leq 1$ and the dominating term on the right-side of (25) is k/m.

Proof. Since $Z_{jl} \sim \text{subG}(1/m)$ for any $j \in [n]$ and $l \in [m]$, we have from Lemma D.2 that $Z_{jl}^2 \sim \text{subexp}(c_1/m^2, c_2/m)$ for some absolute constants c_1 and c_2. Using properties of subexponential random variables from Lemma D.3, we can show that the normalized sum $\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^2$ is also subexponential with parameters $(c_1/mn, c_2/mn)$. Noting that $\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^2\right] = 1$ and using the tail bound from Lemma D.1 we get for $t > 0$,

$$\Pr\left(\left|\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^2 - 1\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{mn}{c_1}t^2, \frac{mn}{c_2}t\right\}\right). \tag{26}$$

For the right side to be at most $\varepsilon > 0$, we see that it suffices to have

$$t \geq \max\left\{\sqrt{\frac{c_1}{mn}\log\frac{2}{\varepsilon}}, \frac{c_2}{mn}\log\frac{2}{\varepsilon}\right\}.$$

Substituting the above into (26) gives us the result. □

Lemma B.2. *Let $Z_1, \ldots, Z_n$ be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Then, there exist absolute constants c_1 and c_2 such that for any $\varepsilon > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^4 \leq \min\left\{\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{1}{\varepsilon}}\right)^2, \left(1 - \frac{c_2}{mn}\log\frac{1}{\varepsilon}\right)^2\right\}\right) \leq \varepsilon.$$

Proof. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^4\right] = 1 + 2/m$, and $t < \mu$. Then, using Jensen's inequality, we have

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^4 \leq \mu - t\right) &\leq \Pr\left(\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^2 \leq \sqrt{\mu - t}\right) \\ &= \Pr\left(\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^2 - 1 \leq -t'\right),\end{aligned}$$

where $t' = 1 - \sqrt{\mu - t}$.

We can now use Lemma B.1 to get

$$\Pr\left(\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^2 - 1 \leq -t'\right) \leq \exp\left(-\min\left\{\frac{mn}{c_1}(t')^2, \frac{mn}{c_2}t'\right\}\right).$$

Substituting for t', we get

$$\Pr\left(\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^4 \leq \mu - t\right) \leq \exp\left(-\min\left\{\frac{mn}{c_1}(1 - \sqrt{\mu - t})^2, \frac{mn}{c_2}(1 - \sqrt{\mu - t})\right\}\right).$$

Equating the expression on the right to $\varepsilon > 0$, the inequality above can be rewritten as

$$\Pr\left(\frac{1}{n}\sum_{j=1}^n \|Z_j\|_2^4 \leq \min\left\{\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{1}{\varepsilon}}\right)^2, \left(1 - \frac{c_2}{mn}\log\frac{1}{\varepsilon}\right)^2\right\}\right) \leq \varepsilon.$$

□

Lemma B.3. *Let $Z_1, \ldots, Z_n$ be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian* *entries with variance $1/m$. Then, there exist absolute constants c_1 and c_2 such that for any $\varepsilon > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^4 \geq \mathit{max}\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{n}{\varepsilon}}\right)^2, \left(1+\frac{c_2}{m}\log\frac{n}{\varepsilon}\right)^2\right\}\right) \leq \varepsilon.$$

Proof. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^4\right]$ and note as we did in Lemma B.1 that $\|Z_j\|_2^2 \sim \text{subexp}(c_1/m, c_2/m)$. We have by union bound that

$$\begin{aligned}
\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^4 \geq \mu + t\right) &\leq \sum_{j=1}^{n}\Pr\left(\|Z_j\|_2^4 \geq \mu + t\right) \\
&= \sum_{j=1}^{n}\Pr\left(\|Z_j\|_2^2 - 1 \geq \sqrt{\mu + t} - 1\right) \\
&\leq n\exp\left(-\min\left\{\frac{m(t')^2}{c_1}, \frac{mt'}{c_2}\right\}\right)
\end{aligned}$$

where the last inequality follows from Lemma D.1 with $t' = \sqrt{\mu + t} - 1$. For this probability to be at most ε, we see after some simplification that it suffices to have

$$\mu + t = \max\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{n}{\varepsilon}}\right)^2, \left(1+\frac{c_2}{m}\log\frac{n}{\varepsilon}\right)^2\right\},$$

which gives us the result. □

Lemma B.4. *Let $Z_j, Y_{j1}, \ldots, Y_{j,k-1}$, $j \in [n]$, be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2\right]$. Then, there exist absolute constants c_1 and c_2 such that for any $\varepsilon > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \leq \mu - t\right) \leq \varepsilon,$$

for

$$\begin{aligned}
\mu - t = \frac{k-1}{m}&\left(1-\sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}\right) \\
&- \sqrt{\frac{1}{mn}\log\frac{4}{\varepsilon}}\,\mathit{max}\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\mathit{max}\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}\right), \left(1+\frac{c_2}{m}\log\frac{4n}{\varepsilon}\right)\right\},
\end{aligned}$$

when $n \geq (c_2^2/c_1)\log(4/\varepsilon)$.

Proof. Note that conditioned on Z_j, the random variable $Y_{jl}^\top Z_j$ is subgaussian with parameter $\|Z_j\|_2^2/m$, for any $j \in [n]$ and $l \in [k-1]$. Using Lemmas D.2 and D.3, we have that the normalized sum $(1/n)\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2$, conditioned on $\{Z_j\}_{j=1}^{n}$, is subexponential with parameters v^2 and b where

$$v^2 = \frac{c_1}{m^2n^2}(k-1)\sum_{j=1}^{n}\|Z_j\|_2^4, \quad b = \frac{c_2}{mn}\max_{j\in[n]}\|Z_j\|_2^2.$$

Let $\mu' = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \middle| \{Z_j\}_{j=1}^{n}\right]$. As noted before, $Y_{jl}^\top Z_j \sim \text{subG}(\|Z_j\|_2^2/m)$. Additionally, the variance and the variance parameter are equal, which gives

$$\mu' = \frac{k-1}{mn}\sum_{j=1}^{n}\|Z_j\|_2^2.$$

From Lemma D.1 we have that for $t > 0$,

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^{\top}Z_j)^2 \le \mu - t \middle| \{Z_j\}_{j=1}^n\right) \\
&= \Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^{\top}Z_j)^2 - \mu' \le \mu - t - \mu' \middle| \{Z_j\}_{j=1}^n\right) \\
&\le \exp\left(-\min\left\{\frac{m^2n^2(t')^2}{c_1(k-1)\sum_{j=1}^n\|Z_j\|_2^4}, \frac{mnt'}{c_2\max_{j\in[n]}\|Z_j\|_2^2}\right\}\right) \qquad (27)
\end{aligned}$$

where $t' = \mu' + t - \mu$. We now handle the Z_j-dependent terms in the exponent. In particular, we require upper bounds on the terms in the denominator and a lower bound on μ' that hold with high probability. Recall that from Lemma B.3, we have

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^4 \le \max\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{n}{\varepsilon}}\right)^2, \left(1+\frac{c_2}{m}\log\frac{n}{\varepsilon}\right)^2\right\}\right) \ge 1-\varepsilon.$$

Also, from Lemma B.1, we have that

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2 \ge 1 - \max\left\{\sqrt{\frac{c_1}{mn}\log\frac{1}{\varepsilon}}, \frac{c_2}{mn}\log\frac{1}{\varepsilon}\right\}\right) \ge 1-\varepsilon.$$

Finally, by independence of Z_js,

$$\begin{aligned}
\Pr\left(\max_{j\in[n]}\|Z_j\|_2^2 \le \mu + t\right) &= \prod_{j=1}^{n}\Pr\left(\|Z_j\|_2^2 \le \mu + t\right) \\
&\ge \left(1 - \exp\left(-\min\left\{\frac{m(\mu+t-1)^2}{c_1}, \frac{m(\mu+t-1)}{c_2}\right\}\right)\right)^n \\
&\ge 1 - n\exp\left(-\min\left\{\frac{m(\mu+t-1)^2}{c_1}, \frac{m(\mu+t-1)}{c_2}\right\}\right),
\end{aligned}$$

which gives

$$\Pr\left(\max_{j\in[n]}\|Z_j\|_2^2 \le 1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{n}{\varepsilon}}, \frac{c_2}{m}\log\frac{n}{\varepsilon}\right\}\right) \ge 1-\varepsilon.$$

Using these results together with (27), we have

$$\begin{aligned}
\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^{\top}Z_j)^2 \le \mu - t\right) &\le \exp\left(-\min\left\{\frac{m^2n(\frac{k-1}{m}\beta_1 + t - \mu)^2}{c_1(k-1)\beta_2}, \frac{mn(\frac{k-1}{m}\beta_1 + t - \mu)}{c_2\beta_3}\right\}\right) \\
&\quad + \frac{3\varepsilon}{4}, \qquad (28)
\end{aligned}$$

where

$$\begin{aligned}
\beta_1 &= 1 - \max\left\{\sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}, \frac{c_2}{mn}\log\frac{4}{\varepsilon}\right\}, \\
\beta_2 &= \max\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}\right)^2, \left(1+\frac{c_2}{m}\log\frac{4n}{\varepsilon}\right)^2\right\}, \\
\beta_3 &= 1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}, \frac{c_2}{m}\log\frac{4n}{\varepsilon}\right\}.
\end{aligned}$$

Now, the first term on the right side of (28) equals $\varepsilon/4$ if

$$\mu - t = \frac{k-1}{m}\beta_1 - \max\left\{\sqrt{\frac{c_1\beta_2(k-1)}{m^2 n}}\log\frac{4}{\varepsilon}, \frac{c_2\beta_3}{mn}\log\frac{4}{\varepsilon}\right\}. \tag{29}$$

The expression above can be simplified under some mild assumptions on n. In particular, when $mn \geq (c_2^2/c_1)\log(4/\varepsilon)$ and $m \geq (c_2^2/c_1)\log(4n/\varepsilon)$, then (29) simplifies to

$$\mu - t = \frac{k-1}{m}\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}\right) - \sqrt{\frac{1}{mn}\log\frac{4}{\varepsilon}}\max\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\left(1 + \sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}\right).$$

On the other hand, when $mn \geq (c_2^2/c_1)\log 4/\varepsilon$ and $m < (c_2/\sqrt{c_1})\log(4n/\varepsilon)$, we have

$$\mu - t = \frac{k-1}{m}\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}\right) - \sqrt{\frac{1}{mn}\log\frac{4}{\varepsilon}}\max\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\left(1 + \frac{c_2}{m}\log\frac{4n}{\varepsilon}\right),$$

which gives us the following simplified version of (29) when $n \geq (c_2^2/c_1)\log(4/\varepsilon)$:

$$\begin{aligned}\mu - t = {} & \frac{k-1}{m}\left(1 - \sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}\right) \\ & - \sqrt{\frac{1}{mn}\log\frac{4}{\varepsilon}}\max\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\max\left\{\left(1 + \sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}\right), \left(1 + \frac{c_2}{m}\log\frac{4n}{\varepsilon}\right)\right\}.\end{aligned}$$

This completes the proof. □

Lemma B.5. *Let $Z_j, Y_{j1}, \ldots, Y_{j,k-1}$, $j \in [n]$, be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2\right]$. Then, there exist absolute constants c_1 and c_2 such that for any $\varepsilon > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \geq \mu + t\right) \leq \varepsilon,$$

for

$$\begin{aligned}\mu + t = {} & \frac{k-1}{m}\left(1 + \sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}\right) \\ & + \sqrt{\frac{1}{mn}\log\frac{4}{\varepsilon}}\, \mathit{max}\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\mathit{max}\left\{\left(1 + \sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}\right), \left(1 + \frac{c_2}{m}\frac{4n}{\varepsilon}\right)\right\},\end{aligned}$$

when $n \geq (c_2^2/c_1)\log(4/\varepsilon)$.

Proof. The proof is similar to that of Lemma B.4. We start by noting that conditioned on $\{Z_j\}_{j=1}^n$, the normalized sum $(1/n)\sum_{j=1}^n\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2$ is subexponential with parameters v^2 and b where

$$v^2 = \frac{c_1}{m^2 n^2}(k-1)\sum_{j=1}^{n}\|Z_j\|_2^4, \quad b = \frac{c_2}{mn}\max_{j\in[n]}\|Z_j\|_2^2.$$

Again, using the tail bound for subexponential random variables, we get

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \geq \mu + t \,\middle|\, \{Z_j\}_{j=1}^n\right) &= \Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 - \mu' \geq \mu + t - \mu' \,\middle|\, \{Z_j\}_{j=1}^n\right) \\ &\leq \exp\left(-\min\left\{\frac{m^2n^2(t')^2}{c_1(k-1)\sum_{j=1}^n\|Z_j\|_2^4}, \frac{mnt'}{c_2\max_{j\in[n]}\|Z_j\|_2^2}\right\}\right)\end{aligned} \tag{30}$$

where

$$\mu' = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \middle| \{Z_j\}_{j=1}^n\right] = \frac{k-1}{mn}\sum_{j=1}^{n}\|Z_j\|_2^2,$$

and $t' = \mu + t - \mu'$. To handle the Z_j-dependent terms in the exponent, we require high probability upper bounds on the terms in the denominator and on μ'. Proceeding as in the proof of Lemma B.4, we have the following bounds on the terms in the denominator in (30):

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^4 \le \max\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{n}{\varepsilon}}\right)^2, \left(1+\frac{c_2}{m}\log\frac{n}{\varepsilon}\right)^2\right\}\right) \ge 1-\varepsilon.$$

and

$$\Pr\left(\max_{j\in[n]}\|Z_j\|_2^2 \le 1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{n}{\varepsilon}}, \frac{c_2}{m}\log\frac{n}{\varepsilon}\right\}\right) \ge 1-\varepsilon. \tag{31}$$

Also, from Lemma B.1,

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2 \le 1+\max\left\{\sqrt{\frac{c_1}{mn}\log\frac{1}{\varepsilon}}, \frac{c_2}{mn}\log\frac{1}{\varepsilon}\right\}\right) \ge 1-\varepsilon. \tag{32}$$

We note that although a high probability upper bound on $\max_{j\in[n]}\|Z_j\|_2^2$ implies a high probability upper bound on $(1/n)\sum_{j=1}^n\|Z_j\|_2^2$, we specifically use the bound in (32) since the deviation term has better dependence on n (which is lost in (31) due to a union bound step).

Using these results along with (30), we have

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \ge \mu+t\right) \le \exp\left(-\min\left\{\frac{m^2 n(\mu+t-\frac{k-1}{m}\beta_1)^2}{c_1(k-1)\beta_2}, \frac{mn(\mu+t-\frac{k-1}{m}\beta_1)}{c_2\beta_3}\right\}\right) + \frac{3\varepsilon}{4}, \tag{33}$$

where

$$\beta_1 = 1+\max\left\{\sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}, \frac{c_2}{mn}\log\frac{4}{\varepsilon}\right\},$$

$$\beta_2 = \max\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}\right)^2, \left(1+\frac{c_2}{m}\log\frac{4n}{\varepsilon}\right)^2\right\},$$

and

$$\beta_3 = 1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}, \frac{c_2}{m}\log\frac{4n}{\varepsilon}\right\} = \sqrt{\beta_2}.$$

Simplifying as we did in Lemma B.4 under the assumption that $n \ge (c_2^2/c_1)\log(4/\varepsilon)$, we see that if

$$\begin{aligned}\mu+t =& \frac{k-1}{m}\left(1+\sqrt{\frac{c_1}{mn}\log\frac{4}{\varepsilon}}\right)\\ &+\sqrt{\frac{1}{mn}\log\frac{4}{\varepsilon}}\max\left\{\sqrt{c_1\frac{k-1}{m}}, c_2\right\}\max\left\{\left(1+\sqrt{\frac{c_1}{m}\log\frac{4n}{\varepsilon}}\right), \left(1+\frac{c_2}{m}\frac{4n}{\varepsilon}\right)\right\},\end{aligned}$$

then the first term on the right side of (33) is less than $\varepsilon/4$, which completes the proof. □

APPENDIX C
FOURTH MOMENT OF THE MINIMUM EIGENVALUE OF A WISHART MATRIX

Lemma C.1. *Let $\Phi \in \mathbb{R}^{m\times k}$ with independent $\mathcal{N}(0,1)$ entries and let $A = \Phi\Phi^\top$. If Z denotes the minimum eigenvalue of A, then for $k-m>7$,*

$$\mathbb{E}\left[Z^{-4}\right] \leq \frac{c}{k^4(1-m/k)^8}.$$

Proof. Since Z is a nonnegative random variable, we have that

$$\begin{aligned}\mathbb{E}\left[Z^{-4}\right] &= \int_0^\infty \Pr\left(Z^{-4} \geq u\right) du \\ &= \int_0^\theta \Pr\left(Z^{-4} \geq u\right) du + \int_\theta^\infty \Pr\left(Z^{-4} \geq u\right) du \\ &\leq \theta + \int_\theta^\infty \Pr\left(Z \leq u^{-\frac{1}{4}}\right) du,\end{aligned}$$

for some $\theta > 0$. Substituting $u^{-\frac{1}{4}} = k\varepsilon^2$, we get

$$\mathbb{E}\left[Z^{-4}\right] \leq \theta + \frac{8}{k^4}\int_0^{\frac{\theta^{-\frac{1}{8}}}{\sqrt{k}}} \Pr\left(Z \leq k\varepsilon^2\right)\frac{1}{\varepsilon^9}d\varepsilon.$$

The density of the smallest eigenvalue of a Wishart matrix with parameters k and m (A in this case) is known in closed form [9, Lemma 4.1], which we restate here:

$$\begin{aligned}\Pr\left(Z \leq k\varepsilon^2\right) &\leq \frac{1}{\Gamma(k-m+2)}(\varepsilon k)^{k-m+1} \\ &\leq \left(\frac{e}{k-m+1}\right)^{k-m+1}(\varepsilon k)^{k-m+1},\end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function and $\Gamma(n) = (n-1)!$ for integer n. Using this, we get

$$\begin{aligned}\mathbb{E}Z^{-4} &\leq \theta + \frac{8}{k^4}\left(\frac{ek}{k-m+1}\right)^{k-m+1}\int_0^{\frac{\theta^{-\frac{1}{8}}}{\sqrt{k}}}(\varepsilon)^{k-m-8}d\varepsilon \\ &= \theta + \frac{8}{k^4}\left(\frac{ek}{k-m+1}\right)^{k-m+1}\frac{1}{k-m-7}\left(\frac{\theta^{-\frac{1}{4}}}{k}\right)^{\frac{k-m-7}{2}}.\end{aligned}$$

Choosing $\theta = \left(\frac{e\sqrt{k}}{k-m+1}\right)^8$ and simplifying gives

$$\mathbb{E}\left[Z^{-4}\right] \leq \frac{9e^8k^4}{(k-m-7)^8} \leq \frac{c}{k^4\left(1-\frac{m}{k}\right)^8}.$$

□

APPENDIX D
PRELIMINARIES

Lemma D.1. *Let X be a subexponential random variable with parameters v^2 and $b>0$ (denoted $X \sim$ subexp(v^2,b)), that is,*

$$\mathbb{E}\left[\exp(\theta(X-\mathbb{E}\left[X\right]))\right] \leq \exp\left(\frac{\theta^2v^2}{2}\right), \quad |\theta| < \frac{1}{b}.$$

Then,

$$\Pr\left(|X - \mathbb{E}[X]| \geq t\right) \leq 2\exp\left(-\mathit{min}\left\{\frac{t^2}{2v^2}, \frac{t}{2b}\right\}\right).$$

Proof. See [32, Proposition 2.2]. ☐

The next lemma is a standard result on the relation between subgaussian and subexponential random variables. We provide the proof for the specific form we need.

Lemma D.2. *Let $X \sim \mathit{subG}(\sigma^2)$ with $\mathbb{E}[X] = 0$. Then $X^2 \sim \mathit{subexp}(128\sigma^4, 8\sigma^2)$.*

Proof. Let $Y = X^2$. We start by upper bounding the moment generating function (MGF) of Y. For $\theta > 0$,

$$\begin{aligned}
\mathbb{E}\left[e^{\theta(Y-\mathbb{E}[Y])}\right] &= \mathbb{E}\left[\sum_{q=0}^{\infty}\frac{(\theta(Y-\mathbb{E}[Y]))^q}{q!}\right] \\
&\leq 1 + \sum_{q=2}^{\infty}\frac{(2\theta)^q\mathbb{E}[Y^q]}{q!} \\
&= 1 + \sum_{q=2}^{\infty}\frac{(2\theta)^q}{q!}\mathbb{E}\left[X^{2q}\right],
\end{aligned}$$

where in the second step we used $(\mathbb{E}[|Y-\mathbb{E}[Y]|^q])^{\frac{1}{q}} \leq (\mathbb{E}[|Y|^q])^{\frac{1}{q}} + \mu \leq 2(\mathbb{E}[Y^q])^{\frac{1}{q}}$.

Now, for $X \sim \text{subG}(\sigma^2)$, we have the following upper bound on the moments of X from [6, Theorem 2.1] :

$$\mathbb{E}\left[X^{2q}\right] \leq 2q!2^q\sigma^{2q}.$$

This gives

$$\begin{aligned}
\mathbb{E}\left[e^{\theta(Y-\mathbb{E}[Y])}\right] &\leq 1 + 2\sum_{q=2}^{\infty}\frac{\theta^q}{q!}q!2^{2q}\sigma^{2q} \\
&= 1 + \frac{32\theta^2\sigma^4}{1-4\theta\sigma^2}, \quad \theta < \frac{1}{4\sigma^2}.
\end{aligned}$$

For $\theta \leq 1/8\sigma^2$, we get

$$\begin{aligned}
\mathbb{E}\left[e^{\theta(Y-\mathbb{E}[Y])}\right] &\leq 1 + 64\theta^2\sigma^4 \\
&\leq e^{64\theta^2\sigma^4},
\end{aligned}$$

that is, $Y \sim \text{subexp}(128\sigma^4, 8\sigma^2)$. ☐

Lemma D.3. *Let $X_i \sim \mathit{subexp}(v_i^2, b_i)$ be independent subexponential random variables for $i \in [n]$. Then, for a constant $a \in \mathbb{R}$, we have that $aX_1 \sim \mathit{subexp}(a^2v_1^2, |a|b_1)$ and $\sum_{i=1}^n X_i \sim \mathit{subexp}(\sum_{i=1}^n v_i^2, \mathit{max}_{i\in[n]}b_i)$.*

Proof. The proof involves bounding the MGF of the transformed random variables and noting that it has the same form as the MGF of a subexponential random variable with the parameters appropriately transformed. Specifically, for $0 < \theta < 1/|a|b_1$, we have

$$\mathbb{E}\left[\exp(a\theta(X_1 - \mathbb{E}[X_1]))\right] \leq \exp\left(\frac{a^2\theta^2v_1^2}{2}\right),$$

that is, $aX_1 \sim \text{subexp}(a^2v_1^2, |a|b_1)$.

Similarly, bounding the MGF of the sum $Y = \sum_{i=1}^{n} X_i$, we get

$$\begin{aligned}\mathbb{E}\left[\exp(\theta(Y - \mathbb{E}\left[Y\right]))\right] &= \prod_{i=1}^{n} \mathbb{E}\left[\exp(\theta(X_i - \mathbb{E}\left[X_i\right]))\right] \\ &\leq \prod_{i=1}^{n} \exp\left(\frac{\theta^2 v_i^2}{2}\right),\end{aligned}$$

when $|\theta| < 1/b_i$ for all $i \in [n]$. That is, for $|\theta| < 1/(\max_{i\in[n]} b_i)$,

$$\mathbb{E}\left[\exp(\theta(Y - \mathbb{E}\left[Y\right]))\right] \leq \exp\left(\theta^2 \sum_{i=1}^{n} \frac{v_i^2}{2}\right)$$

which shows that $Y \sim \text{subexp}(\sum_{i=1}^{n} v_i^2, \max_{i\in[n]} b_i)$. □

Lemma D.4. *Let W and Z be m-dimensional random vectors having independent zero-mean subgaussian entries with variance $1/m$ and fourth moment $3/m^2$. Then,*

$$\mathbb{E}\left[\|Z\|_2^4\right] = 1 + \frac{2}{m}, \quad \textit{and} \quad \mathbb{E}\left[(Z^\top W)^2\right] = \frac{1}{m}.$$

Proof. The proof is based on a straightforward calculation. We have

$$\begin{aligned}\mathbb{E}\left[\|Z\|_2^4\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{m} Z_i^2\right)^2\right] \\ &= \sum_{i=1}^{m} \mathbb{E}\left[Z_i^4\right] + \sum_{i\neq j} \mathbb{E}\left[Z_i^2 Z_j^2\right] \\ &= m\frac{3}{m^2} + m(m-1)\frac{1}{m^2} \\ &= 1 + \frac{2}{m},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[(Z^\top W)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(Z^\top W)^2 | Z\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{i=1}^{m} Z_i^2 W_i^2 + \sum_{i\neq j} Z_i W_i Z_j W_j\right)\middle| Z\right]\right] \\ &= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i^2\right] \\ &= \frac{1}{m}.\end{aligned}$$

□

Lemma D.5. *Let $A, B \in \mathbb{R}^{m\times m}$ be symmetric, positive definite matrices and let $a_1 \geq \cdots \geq a_m$ and $b_1 \geq \cdots \geq b_m$ denote their respective ordered eigenvalues. Then,*

$$\textit{Tr}(AB) \leq \sum_{i=1}^{m} a_i b_i.$$

Proof. Let $\gamma_1, \cdots, \gamma_m$ and $s_1 \geq \cdots \geq s_m$ denote the eigenvalues and singular values of AB, respectively. Note that γ_i's can be complex in general since AB need not be symmetric. We start by noting that

$$\text{Tr}(AB) = \sum_{i=1}^{m} \gamma_i \leq \sum_{i=1}^{m} |\gamma_i| \leq \sum_{i=1}^{m} s_i, \tag{34}$$

where the last inequality follows from [16] [Theorem 3.3.13]. The next step is to relate the sum of the singular values of AB to the eigenvalues of A and B. We use the following two results from [16] [Theorem 3.3.4, Corollary 3.3.10]:

(i) the product of singular values of AB can be upper bounded as

$$\prod_{i=1}^{m} s_i \leq \prod_{i=1}^{m} a_i b_i;$$

(ii) for nonnegative real numbers $\alpha_1 \geq \cdots \geq \alpha_m$ and $\beta_1 \geq \cdots \geq \beta_m$, if

$$\prod_{i=1}^{m} \alpha_i \leq \prod_{i=1}^{m} \beta_i,$$

then

$$\sum_{i=1}^{m} \alpha_i \leq \sum_{i=1}^{m} \beta_i.$$

From the results above, we have that

$$\sum_{i=1}^{m} s_i \leq \sum_{i=1}^{m} a_i b_i,$$

which together with (34) gives the result. □

ACKNOWLEDGMENT

The authors would like to thank Manjunath Krishnapur for pointing them to useful references on eigenvalues of Wishart random matrices.

REFERENCES

[1] J. Acharya, C. Canonne, and H. Tyagi, "Inference under information constraints I: Lower bounds from chi-square contraction," *CoRR*, vol. abs/1812.11476, 2018.

[2] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Trans. on Inf. Theory*, vol. 56, no. 10, pp. 5111–5130, 2010.

[3] M. Azizyan, A. Krishnamurthy, and A. Singh, "Extreme compressive sampling for covariance estimation," *IEEE Trans. Inf. Theory*, vol. 64, no. 12, pp. 7613–7635, Dec. 2018.

[4] O. Balkan, K. Kreutz-Delgado, and S. Makeig, "Localization of more sources than sensors via jointly-sparse Bayesian learning," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 131–134, 2014.

[5] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265 – 274, 2009.

[6] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[7] T. T. Cai and A. Zhang, "ROP: Matrix recovery via rank-one projections," *The Annal of Statistics*, vol. 43, no. 1, pp. 102–138, Feb 2015.

[8] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.

[9] Z. Chen and J. J. Dongarra, "Condition numbers of Gaussian random matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 3, pp. 603–620, 2005.

[10] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[11] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.

[12] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. on Inf. Theory*, vol. 55, no. 12, pp. 5758–5772, 2009.

[13] S. Foucart, "Recovering jointly sparse vectors via hard thresholding pursuit," in *Sampling Theory and Applications (SAMPTA), Singapore*, 2011.

[14] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 655–687, Dec 2008.

[15] A. J. Hoffman and H. W. Wielandt, "The variation of the spectrum of a normal matrix," *Duke Math. J.*, vol. 20, no. 1, pp. 37–39, 03 1953.

[16] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[17] Y. Jin and B. D. Rao, "Support recovery of sparse signals in the presence of multiple measurement vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3139–3157, May 2013.

[18] S. Khanna and C. R. Murthy, "Rényi divergence based covariance matching pursuit of joint sparse support," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications, (SPAWC), Sapporo, Japan*, 2017, pp. 1–5.

[19] A. Koochakzadeh, H. Qiao, and P. Pal, "On fundamental limits of joint sparse support recovery using certain correlation priors," *IEEE Trans. Signal Process.*, vol. 66, no. 17, pp. 4612–4625, Sep. 2018.

[20] D. M. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8-2, pp. 3010–3022, 2005.

[21] P. Pal and P. P. Vaidyanathan, "Pushing the limits of sparse support recovery using correlation information," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 711–726, 2015.

[22] S. Park, N. Y. Yu, and H. Lee, "An information-theoretic study for joint sparsity pattern recovery with different sensing matrices," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5559–5571, Sep. 2017.

[23] L. Ramesh and C. R. Murthy, "Sparse support recovery via covariance estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada*, 2018, pp. 6633–6637.

[24] G. Reeves and M. Gastpar, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3451–3465, 2013.

[25] J. Scarlett and V. Cevher, "Limits on support recovery with probabilistic models: An information-theoretic framework," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 593–620, 2017.

[26] A. Sharma and C. R. Murthy, "Group testing-based spectrum hole search for cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3794–3805, 2014.

[27] G. Tang and A. Nehorai, "Performance analysis for sparse support recovery," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1383–1399, 2010.

[28] J. A. Tropp, "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, Mar. 2006.

[29] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *IEEE International Conference on Acoustics, Speech, and signal Processing (ICASSP), Philadelphia, PA, USA*, Mar. 2005, pp. 721–724.

[30] ——, "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, Mar. 2006.

[31] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, 2009.

[32] ——, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[33] C. Wang, Q. Zhao, and C. Chuah, "Group testing under sum observations for heavy hitter detection," in *Information Theory and Applications Workshop (ITA), San Diego, CA, USA*, Feb. 2015, pp. 149–153.

[34] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7-2, pp. 3704–3716, 2007.