Multiple Support Recovery Using Very Few Measurements Per Sample

Lekshmi Ramesh, Chandra R. Murthy, and Himanshu Tyagi

***Abstract*—In the problem of multiple support recovery, we are given access to linear measurements of multiple sparse samples in $\mathbb{R}^d$. These samples can be partitioned into ℓ groups, with samples having the same support belonging to the same group. For a given budget of m measurements per sample, the goal is to recover the ℓ underlying supports, in the absence of the knowledge of group labels. We study this problem with a focus on the *measurement-constrained* regime where m is smaller than the support size k of each sample. We design a two-step procedure that estimates the union of the underlying supports first, and then uses a spectral algorithm to estimate the individual supports. Our proposed estimator can recover the supports with $m < k$ measurements per sample, from $\tilde{O}(k^4\ell^4/m^4)$ samples. Our guarantees hold for a general, generative model assumption on the samples and measurement matrices. We also provide results from experiments conducted on synthetic data and on the MNIST dataset.**

***Index Terms*—Compressed sensing, support recovery, concentration inequalities, spectral clustering.**

I. Introduction

We study the problem of *multiple support recovery* using linear measurements, where there are n random samples $X_1, \ldots, X_n$ taking values in $\mathbb{R}^d$, such that for each $i \in [n]$, $\texttt{supp}(X_i) \in \{\mathcal{S}_1, \ldots, \mathcal{S}_\ell\}$ *almost surely,*[1] with $\mathcal{S}_i \subset [d]$ and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for all $i \neq j$. We assume that the samples X_i are sparse and that $|\mathcal{S}_i| = k \ll d$, $i \in [\ell]$. We are given low dimensional projections of these samples using $m \times d$ matrices $\Phi_1, \ldots, \Phi_n$. In our setting, we focus on the regime where we have access to very few measurements per sample, namely, when $m < k$. Given access to the projections $Y_i = \Phi_i X_i, i \in [n]$, and the projection matrices, we seek to recover the underlying supports $\{\mathcal{S}_1, \ldots, \mathcal{S}_\ell\}$.

This is a generalization of the well-studied problem of recovering a *single* unknown support from multiple linear measurements [1]–[4], which has been applied to solve inverse problems in imaging, source localization, and anomaly detection [5]–[8]. It is also related to the study of sparse random effects in mixed linear models [9], [10]. Mixed linear models are a generalization of linear models where an additional additive correction component is included to model a class-specific correction to the average behavior. This residual correction term is commonly known as the random effect term. It is often assumed to be generated from an unknown prior distribution with zero-mean, coming from a parametric family whose parameters are estimated by using the class-specific data. The problem of multiple support recovery is also discussed in [11], [12] under the assumption of slowly varying supports.

There are two sets of unknowns in the setting described above – the labels, indicating which support was chosen for each sample, and the ℓ supports $\mathcal{S}_1, \ldots, \mathcal{S}_\ell$. Note that given the knowledge of the labels, one could group together samples with the same support, and use standard algorithms to recover the support. However, in the absence of labels, the problem of recovering the supports is much harder. A naive scheme could be to just estimate each support individually, which requires $m = O(k \log(d-k))$ measurements per sample [13], [14]. But can we do better if we exploit the joint structure present across the samples, since there will be several samples that have the same support? In this work, we show that one can operate in the measurement-constrained regime of $m < k$, when a sufficiently large number of samples is available.

A. Prior work

For the special case with $n = \ell = 1$, when there is a single k-sparse sample of length d, it is known that $m = \Theta(k \log(d-k))$ measurements are necessary and sufficient to recover the support [13] with noisy measurements, when the inputs are worst-case. For the case with a single common support across multiple samples (i.e., $\ell = 1$ and $n > 1$), several previous works have studied the question of support recovery in the $m > k$ setting [1]–[3].

On the other hand, in the $m < k$ regime, it was shown recently in [4], [15] that $n = \Theta((k^2/m^2)\log d)$ samples are necessary and sufficient, assuming a subgaussian generative model on the samples and measurement matrices and that the measurement matrices are drawn independently across samples. In fact, the lower bound of [4] applies to the worst-case setting as well, showing that while k overall measurements[2] suffice when m exceeds k, at least (roughly) k^2/m measurements are required when $m < k$.

In [16], the problem of recovering the union of supports from linear measurements is considered. The setting allows for overlaps in the supports, but otherwise places no constraints. The results when applied to the case of disjoint supports lead to a requirement of $m = O(k \log d)$ measurements per sample, and therefore are not applicable to our setting. Another line of related works is on multi-task learning/multi-task sparse estimation [17]–[19] that use hierarchical Bayesian models and

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.

[1]The support of a vector $x \in \mathbb{R}^d$ is the set $\{u \in [d] : x_u \neq 0\}$.

[2]The overall measurements in our model are nm.

focus on recovering the samples, rather than the supports, and so still require at least k measurements per sample. However, none of these results shed light on how to recover multiple supports when we are constrained to observe less than k measurements per sample.

We note that there has been some recent work in the literature on mixture of sparse linear regressions that considers the related problem of recovering multiple sparse vectors from linear measurements [19]–[24]. The model shares some similarities with the $m = 1$ case in our setting, but there are some important differences. Unlike our setting, these works consider the samples to be deterministic and do a worst-case analysis. Further, when $\ell = 1$ in the mixture of sparse linear regressions setting, we have multiple observations from the same unknown sparse vector, thus reducing the problem to the standard compressed sensing problem. On the other hand, with $\ell = m = 1$ in our setting, we obtain a single observation from different sparse vectors sharing a common support. The latter setting is harder and requires $\Omega(k^2 \log d)$ samples to recover the common support [4].

B. Contributions and Techniques

Our approach builds on the following simple but crucial observation: since each sample is k-sparse with support equal to one of the $\mathcal{S}_i$ (with the $\mathcal{S}_i$ being disjoint), the sample covariance matrix $(1/n)\sum_{i=1}^{n} X_i X_i^\top$ exhibits a block structure under an unknown permutation of rows and columns. This motivates the use of spectral clustering to recover the underlying supports. However, we only have access to low-dimensional projections of the data.

To circumvent this difficulty, we compute $\Phi_i^\top Y_i$ and use these as a proxy for the data, and form an estimate of the diagonal entries of the covariance matrix of the samples. We build further on this idea and propose an estimator that first determines the union of the ℓ supports from $\Phi_i^\top Y_i$ using the estimator in [4]. We then construct an affinity matrix using the proxy samples $\Phi_i^\top Y_i$ and apply spectral clustering to estimate individual supports from the union.

This clustering based approach to support recovery is new, and very different from traditional approaches to sparse recovery in the multiple sample setting. It reduces the support recovery problem to that of recovering the structure of a certain block matrix, a question which has been studied in the literature on community detection on graphs [25]–[28], and for which many algorithms are known. However, unlike the community detection problem where an instance of the adjacency matrix is available as an observation, the affinity matrix constructed in our case has a more complicated structure and requires a separate, careful analysis.

We show that using our algorithm, it is possible to recover all the supports with *fewer* than k measurements per sample. Our algorithm is easy to implement and has computational complexity that scales linearly with ambient dimension d and number of samples n. Our main result is an upper bound on the sample complexity of the multiple support recovery problem, stated in Theorem 1. In similar spirit to [4], which studied the case of a single unknown support in the measurement-constrained regime of $m < k$, our work provides an algorithm for the multiple support recovery problem in this regime. The analysis of our algorithm involves studying spectral properties of the (random) affinity matrix that has dependent and heavy-tailed entries. We characterize these spectral quantities for the expected affinity matrix, which we show has a block structure, and then use results from matrix perturbation and matrix concentration to obtain performance guarantees for our algorithm.

Also, we provide experimental results on synthetic and real datasets, and show that the proposed algorithm is able to recover the unknown supports with very few measurements per sample. While our guarantees are for the case of disjoint supports, some simple heuristics can be used to handle the case of overlapping supports in practice, as we show in Section V.

C. Organization

In the next section, we formally state the problem and the assumptions we make in our generative model setting. This is followed by a statement of our main result, which provides an upper bound on the sample complexity of multiple support recovery. We describe the estimator in Section III, and analyze its performance in Section IV. We provide experimental results in Section V. The technical results required for the proofs in Section IV are available in the appendices, and algebraic details of the proofs are provided in the supplementary material [29].

D. Notation

For a matrix A, we denote its (u, v)th entry by A_{uv}. For a collection of matrices $\{A_i\}_{i=1}^n$, we use $A_i(u, v)$ to denote the (u, v)th entry of the ith matrix. Also, for a vector X_j, X_{ji} denotes the ith component of X_j. For sets $\mathcal{S}$ and $\mathcal{S}'$, $\mathcal{S}\Delta\mathcal{S}' = (\mathcal{S}\backslash\mathcal{S}') \cup (\mathcal{S}'\backslash\mathcal{S})$ denotes their symmetric difference. For a vector $a \in \mathbb{R}^d$, $\mathtt{supp}\,(a)$ denotes the subset $\{i \in [d] : a_i \neq 0\}$, $\mathtt{diag}(a)$ denotes the $d \times d$ diagonal matrix with entries of a on the diagonal, and $[d]$ denotes the set $\{1, 2, \ldots, d\}$. For a matrix A, we use $\|A\|_{op} \stackrel{\text{def}}{=} \sup_{\|x\|_2=1} \|Ax\|_2$ to denote the operator norm of A. When A is symmetric, $\|A\|_{op}$ equals the magnitude of the largest eigenvalue of A. We use the shorthand Z_1^n to denote independent and identically distributed random variables $Z_1, \ldots, Z_n$. For $u > 0$, we use $\Gamma(u) \stackrel{\text{def}}{=} \int_0^\infty x^{u-1} e^{-x} dx$ to denote the gamma function.

II. Problem Formulation and Main Result

We consider a Bayesian setup for modeling samples $X_1, \ldots, X_n$ taking values in $\mathbb{R}^d$ with $\mathtt{supp}\,(X_i) \stackrel{\text{def}}{=} \{j \in [d] : X_{ij} \neq 0\} \in \{\mathcal{S}_1, \ldots, \mathcal{S}_\ell\}$, where $\mathcal{S}_i \subset [d]$ are unknown sets such that $|\mathcal{S}_i| = k$. Specifically, we consider distributions $\mathrm{P}^{(1)}, \ldots, \mathrm{P}^{(\ell)}$ with[3]

$$\mathtt{supp}\left(\mathrm{P}^{(i)}\right) = \{x \in \mathbb{R}^d : \mathtt{supp}(x) = \mathcal{S}_i\}, \quad i \in [\ell],$$

[3]We consider distributions P with densities f_{P} with respect to the Lebesgue measure and define $\mathtt{supp}\,(\mathrm{P}) = \{x \in \mathbb{R}^d : f_{\mathrm{P}}(x) > 0\}$.

and n i.i.d. samples $X_1, \ldots, X_n$ taking values in $\mathbb{R}^d$ and generated from a common mixture distribution

$$\mathrm{P}_{\mathcal{S}_1,\ldots,\mathcal{S}_\ell} = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathrm{P}^{(i)}, \tag{1}$$

parameterized by the tuple $(\mathcal{S}_1 \ldots, \mathcal{S}_\ell)$. In fact, we assume that $\mathrm{P}^{(i)}$ is a multivariate subgaussian distribution (see Appendix B for the definition of a subgaussian random variable) with zero mean and diagonal covariance matrix $K_{\lambda_i} = \mathtt{diag}\,(\lambda_i)$, where the parameter λ_i is a d-dimensional vector for which $\mathtt{supp}\,(\lambda_i) = \mathcal{S}_i$, $i \in [\ell]$. More concretely, we make the following assumption.

Assumption 1. *For a sample $X_j \sim \mathrm{P}^{(i)}$, $j \in [n]$, $i \in [\ell]$, and an absolute constant c, $\mathbb{E}_{\mathrm{P}^{(i)}}\left[X_j X_j^T\right] = \mathtt{diag}\,(\lambda_i)$ with $\lambda_i \in \mathbb{R}_+^d$, $\mathtt{supp}\,(\lambda_i) = \mathcal{S}_i$, and X_j has independent entries with its tth entry X_{jt} satisfying $X_{jt} \sim$ subG$(c\lambda_{it})$, $t \in [d]$. Furthermore, for each $i \in [\ell]$ and $t \in \mathcal{S}_i$, $\lambda_{it} = \lambda_0 > 0$, and $\mathbb{E}_{\mathrm{P}^{(i)}}\left[X_{jt}^4\right] = \rho$.*

For samples $X_1, \ldots, X_n$ generated as above, we are given access to projections $Y_i = \Phi_i X_i$, $i \in [n]$, where the matrices $\Phi_i \in \mathbb{R}^{m\times d}$ are random and independent for different $i \in [n]$. Our analysis requires handling higher order moments of the entries of the measurement matrices, which motivates the following assumption.

Assumption 2. *The $m \times d$ measurement matrices $\Phi_1, \ldots, \Phi_n$ are independent, with entries that are independent and zero-mean. Furthermore, $\Phi_i(u, v) \sim$ subG(c'/m), and the moment conditions $\mathbb{E}\left[\Phi_i(u, v)^2\right] = 1/m$ and $\mathbb{E}\left[\Phi_i(u, v)^{2q}\right] = c_q/m^q$ hold for $q \in \{2, 3, 4\}$, where c_q and c' are absolute constants.*

The assumption above holds, for example, when $\Phi_i(u, v) \sim \mathcal{N}(0, 1/m)$ or when $\Phi_i(u, v)$ are Rademacher, i.e., take values from $\{1/\sqrt{m}, -1/\sqrt{m}\}$ with equal probability. Also, these moment assumptions can be relaxed to hold up to constant factors from above and below, i.e., $\mathbb{E}\left[\Phi_i(u, v)^{2q}\right] = \Theta(1/m^q)$.

Our goal is to recover the supports $\{\mathcal{S}_1, \ldots, \mathcal{S}_\ell\}$ using $\{Y_i, \Phi_i\}_{i=1}^n$. The error criterion will be the average of the per support errors, measured using the set difference between the true and estimated supports. Specifically, denote by $\Sigma'_{\ell,d}$ the set consisting of all ℓ tuples of subsets $(\mathcal{S}_1, \ldots, \mathcal{S}_\ell)$ such that $\mathcal{S}_i \subset [d]$, $i \in [\ell]$, and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$, for all $i \neq j$. Let $\Sigma_{k,\ell,d} \subset \Sigma'_{\ell,d}$ be such that $|\mathcal{S}_i| = k$, for all $i \in [\ell]$. Denote by $\mathcal{G}_\ell \stackrel{\text{def}}{=} \{\sigma : [\ell] \to [\ell]\}$ the set of all permutations on $[\ell]$. We have the following definition.

Definition 1. An (n, ε, δ)*-estimator for* $\Sigma_{k,\ell,d}$ is a mapping $e : (Y_1^n, \Phi_1^n) \mapsto (\hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_\ell) \in \Sigma'_{\ell,d}$ for which

$$\mathrm{P}_{\mathcal{S}_1,\ldots,\mathcal{S}_\ell}\left(\exists\, \sigma \in \mathcal{G}_\ell \text{ s.t. } \frac{1}{\ell}\sum_{i=1}^{\ell}\left|\mathcal{S}_i \Delta \hat{\mathcal{S}}_{\sigma(i)}\right| < k\varepsilon\right) \geq 1 - \delta, \tag{2}$$

for all $(\mathcal{S}_1, \ldots, \mathcal{S}_\ell) \in \Sigma_{k,\ell,d}$, where $\mathcal{S}_1 \Delta \mathcal{S}_2$ denotes the symmetric difference between sets $\mathcal{S}_1$ and $\mathcal{S}_2$.

For fixed $\ell, m, k, d, \varepsilon$, and δ, the least n such that we can find an (n, ε, δ)-estimator for $\Sigma_{k,\ell,d}$ is termed the *sample complexity of multiple support recovery*, which we denote by $n^*(\ell, m, k, d, \varepsilon, \delta)$. In our main result stated below, we provide an upper bound on $n^*(\ell, m, k, d, \varepsilon, \delta)$.

Theorem 1. *Let $m, k, d, \ell \in \mathbb{N}$ with $\log k \geq 2$. Further, let $(\log k\ell)^2 \leq m < k$, and $1/k\ell \leq \varepsilon \leq 1/\ell$. Then, under Assumptions 1 and 2, the sample complexity of multiple support recovery satisfies*

$$n^*(\ell, m, k, d, \varepsilon, \delta) = O\Bigg(\max\left\{\frac{1}{\varepsilon}\left(\frac{k\ell}{m}\right)^4 (\log k)^4 \log k\ell \log\frac{1}{\delta}, \frac{k^2\ell^2}{m^2}\log\frac{k\ell(d-k\ell)}{\delta}\right\}\Bigg).$$

Remark 1. For values of ε lower than $1/\ell k$, the result from Theorem 1 continues to hold with ε set to $1/\ell k$. This is because $\varepsilon = 1/\ell k$ corresponds to exact recovery of the supports.

We present the algorithm that attains this performance in the next section, and prove the theorem in Section IV-C.

Our estimator works in two steps by estimating the union of supports first and then estimating each support, and the sample complexity bound above is obtained by analyzing each of the two steps. To the best of our knowledge, this is the first estimator that can recover multiple supports under the constraint of $m < k$ linear measurements per sample. We also note that for the problem of recovering a single support exactly, it was shown in [4] that roughly $\Omega((k/m)^2 \log k(d-k))$ samples are necessary. Thus, our sample complexity upper bound above matches this lower bound quadratically. However, there is a gap between the lower bound and the upper bound, which is an interesting problem for future research.

III. The Estimator

Our first step will be to recover the union of the ℓ underlying supports, and then refine this estimate to finally recover the individual supports. To estimate the union, we use the estimator described in [15]. Following this, we use a spectral clustering based approach to recover the individual supports. We provide more details in the next two subsections.

A. Recovering the union of supports

We first observe that the samples X_i have an effective covariance matrix whose diagonal has support equal to the union of the supports, which allows us to use the results from [4] to recover the union. Specifically, we form "proxy samples" $\hat{X}_i = \Phi_i^\top Y_i = \Phi_i^\top \Phi_i X_i$ and use the diagonal of the sample covariance matrix of $\hat{X}_i$ as an estimate for the diagonal of the covariance matrix for X_i. We will show that the $k\ell$ largest entries of the recovered diagonal correspond to the union of the supports.

Formally, define $\mathcal{S}_{\text{un}} \stackrel{\text{def}}{=} \cup_{i=1}^{\ell}\mathcal{S}_i$ to be the union of the ℓ unknown disjoint supports and note that $|\mathcal{S}_{\text{un}}| = k\ell$. We use the estimator described in [4] and form the statistic $\tilde{\lambda} \in \mathbb{R}^d$ as follows. First, define vectors $a'_1, \ldots, a'_n$ with entries

$$a'_{ji} \stackrel{\text{def}}{=} (\Phi_{ji}^\top Y_j)^2, \quad i \in [d]. \tag{3}$$

$$\mathbb{E}[T] = \begin{bmatrix} \mu_0 & \mu^s & \mu^d & \mu^d \\ \mu^s & \mu_0 & \mu^d & \mu^d \\ \mu^d & \mu^d & \mu_0 & \mu^s \\ \mu^d & \mu^d & \mu^s & \mu_0 \end{bmatrix} \begin{matrix} \} \mathcal{S}_1 \\ \\ \} \mathcal{S}_2 \\ \\ \end{matrix}$$

Fig. 1: Block structure of the expected clustering matrix when $\ell = 2$ and the supports are disjoint, under appropriate permutation of rows and columns.

Each a'_j, $j \in [n]$, can be thought of as a crude estimate for the variances along the d coordinates obtained using the jth sample. We then define the average of these vectors as

$$\tilde{\lambda} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} a'_j. \tag{4}$$

This statistic captures the variance along each coordinate of X_i. Due to the averaging across samples, we expect a larger value of the statistic along coordinates that are present in at least one of the supports. On the other hand, coordinates that are not present any support should result in a smaller value of the statistic. As shown in [4], such a separation between the estimate values indeed occurs when n is sufficiently large. The algorithm declares the indices of the $k\ell$ largest entries of $\tilde{\lambda}$ as the estimate for $\mathcal{S}_{\text{un}}$. Letting $\tilde{\lambda}_{(1)} \geq \cdots \geq \tilde{\lambda}_{(k\ell)}$ represent the sorted entries of $\tilde{\lambda}$, the estimate $\hat{\mathcal{S}}_{\text{un}}$ for the union is

$$\hat{\mathcal{S}}_{\text{un}} = \{(1), \ldots, (k\ell)\}, \tag{5}$$

where we assume the size of the union to be known. In practice, $\tilde{\lambda}$ can be used to estimate the size of the union as well by sorting the entries of $\tilde{\lambda}$ and using the index where there is a sharp decrease in the values as the estimate for $k\ell$, similar to the approach of using scree plots to determine model order in problems such as PCA [30].

B. Recovering individual supports

We now describe the main step of our algorithm where we partition the coordinates in $\hat{\mathcal{S}}_{\text{un}}$ recovered in the first step into disjoint support estimates $\hat{\mathcal{S}}_1, \ldots, \hat{\mathcal{S}}_\ell$. We will use $a'_1, \ldots, a'_n$ described in (3) for this purpose. Since we now have an estimate for the union, we will restrict a'_i to coordinates in $\hat{\mathcal{S}}_{\text{un}}$, and denote them as $a_i \in \mathbb{R}_+^{k\ell}$. Also, without loss of generality, we set $\hat{\mathcal{S}}_{\text{un}} = [k\ell]$.[4]

Our approach is the following: we construct a $k\ell \times k\ell$ *affinity matrix* T and perform spectral clustering using this matrix, which will partition the coordinates in $[k\ell]$ into ℓ groups. The main step here is to construct an affinity matrix T that can provide reliable clustering, and we will use the per-sample variance estimates $a_1, \ldots, a_n$ for this purpose. The idea is that for any coordinate pair $(u, v) \in [k\ell] \times [k\ell]$, if both u and v belong to the same support, then we expect the product $a_{iu}a_{iv}$ to have a "large" value for most of the sample indices $i \in [n]$. On the other hand, if u and v belong to different supports, then $a_{iu}a_{iv}$ will be close to zero for most $i \in [n]$. Although each a_i individually is not a good estimate for the support of X_i, the averaging over n makes the estimate reliable. Formally, we construct the $k\ell \times k\ell$ matrix T with entries

$$T_{uv} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} a_{ju}a_{jv}, \quad (u, v) \in [k\ell] \times [k\ell]. \tag{6}$$

The key observation here is that the *expected* value of the random matrix T has a block structure when the rows and columns are appropriately permuted, and this block structure corresponds to memberships of each of the indices in $[k\ell]$ to one of the underlying supports. This is illustrated in Figure 1 for $\ell = 2$, and we will examine this structure in detail in the next section. A well-known method to find these memberships is to use spectral clustering [26], [31], which uses properties of the eigenvectors of block-structured matrices to determine the partition. For instance, when $\ell = 2$, the *sign* of the second leading eigenvector of $\mathbb{E}[T]$ provides a way to partition the coordinates in $[k\ell]$ into two groups. When $\ell > 2$, spectral clustering makes use of multiple eigenvectors and a nearest neighbor step to identify the partition. A full description of the solution in the general case is provided in Algorithm 1.

In practice, we only have access to T, and not $\mathbb{E}[T]$ to which the discussion above applies. In what follows, we show that the eigenvectors of T itself suffice, provided we have sufficiently many samples. At a high level, our analysis follows that of spectral clustering in the stochastic block model (SBM) setting and the goal is to show that the eigenvectors of $\mathbb{E}[T]$ and its "perturbed" version T are close to each other. This can be shown using the Davis-Kahan theorem from matrix perturbation theory, which states that the angle between any two corresponding eigenvectors of T and $\mathbb{E}[T]$ is small provided the error matrix $T - \mathbb{E}[T]$ has small spectral norm. The key challenge, therefore, is to control $\|T - \mathbb{E}[T]\|_{op}$.

Unlike typical settings, the entries of T are not independent, in addition to being heavy tailed. Standard methods based on the ε-net argument are, therefore, difficult to apply in this setting. One strategy could be to show exponential concentration around the mean for *each* entry of T. Once each entry of T is bounded with high probability, one can bound the Frobenius norm and therefore the spectral norm of the error matrix. However, the moment generating function (MGF) of each summand in (6) is unbounded, so deriving a tail bound for the sum requires a more careful tail splitting method (see, for example, [32, Exercise 2.1.7]), and leads to measurement matrix dependent quantities that are difficult to handle. Due to the same reason, techniques from matrix concentration that involve bounding the MGF of the summands [33, Theorem 6.1, Theorem 6.2] cannot be used in our setting.

To circumvent this difficulty, we turn to a beautiful result by Rudelson [34], that characterizes the expected value of the quantity $\|T - \mathbb{E}[T]\|_{op}$, when T is a sum of independent rank-one matrices and only requires certain moment assumptions on the summands. This is exactly our setting since (6) can

[4] This is to keep notation simple. For a general $\hat{\mathcal{S}}_{\text{un}}$, we can have a function $g : [k\ell] \to \hat{\mathcal{S}}_{\text{un}}$ that provides the mapping of each coordinate of a_i to its corresponding value in $\hat{\mathcal{S}}_{\text{un}}$ as indicated in step 7 of Algorithm 1.

Algorithm 1: Multiple support recovery

Input: Measurements $\{Y_i\}_{i=1}^n$, Measurement matrices $\{\Phi_i\}_{i=1}^n$, k, ℓ

Output: Support estimates $\hat{\mathcal{S}}_1, \dots, \hat{\mathcal{S}}_\ell$

Form variance estimates $a'_1, \dots, a'_n$ with entries

$$a'_{ji} = (\Phi_{ji}^\top Y_j)^2, \quad i \in [d].$$

Compute

$$\tilde{\lambda} = \frac{1}{n}\sum_{i=1}^n a'_i.$$

Sort entries of $\tilde{\lambda}$ to get $\tilde{\lambda}_{(1)} \geq \cdots \geq \tilde{\lambda}_{(d)}$ and output estimate for union

$$\hat{\mathcal{S}}_{\text{un}} = \{(1), \dots, (k\ell)\}.$$

Restrict $a'_1, \dots, a'_n$ to the coordinates in $\hat{\mathcal{S}}_{\text{un}}$, to get $a_1, \dots, a_n$. Also, let $g : [k\ell] \to \hat{\mathcal{S}}_{\text{un}}$ denote the mapping from the coordinates of a_i to the true coordinate in $\hat{\mathcal{S}}_{\text{un}}$.

Construct affinity matrix $T \in \mathbb{R}^{k\ell \times k\ell}$ as

$$T = \frac{1}{n}\sum_{i=1}^n a_i a_i^\top.$$

Compute the ℓ leading eigenvectors $\hat{v}_1, \dots, \hat{v}_\ell$ of T and let these be the columns of $\hat{V} \in \mathbb{R}^{k\ell \times l}$.

(The ℓ-means step) Find $C = \arg\min_{U \in \mathcal{U}_\ell} \|U - \hat{V}\|_F^2$, where $\mathcal{U}_\ell$ is the set of all $k\ell \times \ell$ matrices with at most ℓ distinct rows.

Denote the indices of identical rows of C as sets $\hat{\mathcal{S}}'_1, \dots, \hat{\mathcal{S}}'_\ell$. Declare

$$\hat{\mathcal{S}}_i = \{g(j) \in \hat{\mathcal{S}}_{\text{un}} : j \in \hat{\mathcal{S}}'_i\}.$$

equivalently be represented as $T = (1/n)\sum_{i=1}^n a_i a_i^\top$. An application of Markov inequality followed by the Davis-Kahan theorem then shows that the eigenvectors of T and $\mathbb{E}[T]$ are close to each other. We provide more details about the analysis in the next section.

IV. Analysis of the estimator

A. Recovering the union: Analysis

Our analysis of the probability of exactly recovering $\mathcal{S}_{\text{un}}$ using the estimator in (5) follows the approach in [4]. The key difference is that the samples are now drawn from a *mixture* of subgaussian distributions. In the next result, we show that if X is drawn from the mixture described in (1), then it is subgaussian with covariance matrix $K_{\lambda_{\text{un}}}$ where $\lambda_{\text{un}} = \lambda_1 \vee \cdots \vee \lambda_\ell$, where $\vee$ denotes entrywise maximum. This helps us to determine the effective parameter that characterizes the input distribution, after which we can use the result from [4]. We prove this result for the two component mixture; it can be extended easily to the general case.

Lemma 2. *Let X and Y be zero-mean subgaussian random variables with parameters a^2 and b^2, respectively. Further, let P_X and P_Y denote the distributions of X and Y. Then, the random variable Z with distribution given by the mixture $q\mathrm{P}_X + (1-q)\mathrm{P}_Y$ with $q \in [0,1]$ is subgaussian with parameter $\max\{a^2, b^2\}$.*

Proof. Upon bounding the MGF of Z, we see that

$$\begin{aligned}\mathbb{E}\left[e^{\theta Z}\right] &= q\mathbb{E}\left[e^{\theta X}\right] + (1-q)\mathbb{E}\left[e^{\theta Y}\right] \\ &\leq q e^{\frac{\theta^2 a^2}{2}} + (1-q)e^{\frac{\theta^2 b^2}{2}} \\ &\leq e^{\frac{\theta^2 c^2}{2}},\end{aligned}$$

where $c = \max\{a, b\}$. □

Thus, the samples $X_1, X_2, \dots, X_n$ have entries that are independent and subgaussian with covariance matrix $K_{\lambda_{\text{un}}}$, where $\lambda_{\text{un}} = \lambda_1 \vee \cdots \vee \lambda_\ell$. Therefore, results from [4] imply that we can recover $\mathcal{S}_{\text{un}}$ from the variance estimate (4) by retaining the $k\ell$ largest entries. In particular, a direct application of [4, Theorem 3] with support size set to $k\ell$, gives us the following result.

Theorem 3. *Let $\hat{\mathcal{S}}_{un}$ described in (5) be the estimate for the union $\mathcal{S}_{un}$. Then, for every $\delta > 0$,*

$$\Pr\left(\hat{\mathcal{S}}_{un} \neq \mathcal{S}_{un}\right) \leq \delta,$$

provided $m \geq (\log k\ell)^2 > 1$, and

$$n \geq c\left(\frac{k^2\ell^2}{m^2}\log\frac{k\ell(d-k\ell)}{\delta}\right),$$

for an absolute constant c.

As we discussed in the introduction, if we had labels for each sample indicating which support it belongs to, we could directly use the estimator from [4] after grouping the samples with the same support together. This would require $O((k^2\ell/m^2)\log k(d-k))$ samples. On the other hand, when the labels are unknown, the number of samples required even to estimate the union of the supports is higher, as seen from the theorem above.

B. Recovering individual supports: Analysis

Our analysis is based on the fact that the expected affinity matrix has a block structure (under an appropriate permutation of its rows and columns), which we prove in the next lemma.

Lemma 4 (Block structure of $\mathbb{E}[T]$)**.** *Under Assumptions 1 and 2, for the matrix $T \in \mathbb{R}^{k\ell \times k\ell}$ in (6), $\mathbb{E}[T]$ has entries given by*

$$\mathbb{E}[T_{uv}] = \begin{cases}\mu_0, & \text{if } u = v, \\ \mu_s, & \text{if } u \neq v, (u,v) \in \mathcal{S}_i \times \mathcal{S}_i \text{ for any } i \in [\ell], \\ \mu_d, & \text{otherwise},\end{cases}$$

where the parameters μ_0, μ_s, and μ_d depend on k, m, and ℓ and can be explicitly calculated.

The proof of Lemma 4 appears in the supplementary material [29] and involves computing the expected values

of expressions containing higher order terms in Φ_i and X_i. Before we proceed, we note the following extension of the "median trick" (see, for example, [35]) which shows that the dependence of sample complexity on δ is at most a factor of $O(\log 1/\delta)$, provided we can find an $(n, \varepsilon, 1/4)$-estimator.

Lemma 5 (Probability of error boosting)**.** *For $\delta \in (0,1)$ and $\ell \in \mathbb{N}$, if we can find an $(n, \varepsilon, 1/4)$-estimator for $\Sigma_{k,\ell,d}$, then we can find an $\left(n\lceil 8 \log \frac{1}{\delta} \rceil, 3\varepsilon, \delta\right)$-estimator for $\Sigma_{k,\ell,d}$.*

We provide the proof in Appendix A-A.

Thus, from here on, we fix our error requirement to $\delta = 1/4$ and seek $(n, \varepsilon, 1/4)$-estimators with the least possible n. We characterize the performance of the clustering step in the following theorem. The analysis of this step is conditioned on exact recovery of the union $\mathcal{S}_{\text{un}}$ in the first step.

Theorem 6. *Let $\nu_1 \geq \cdots \geq \nu_{k\ell}$ denote the ordered eigenvalues of $\mathbb{E}[T] \in \mathbb{R}^{k\ell \times k\ell}$, and define $\Delta_\ell = \nu_\ell - \nu_{\ell+1}$ when $\ell \geq 2$. For every $\varepsilon \in [1/\ell k, 1/\ell)$, we can find an $(n, \varepsilon, 1/4)$-estimator for $\Sigma_{k,\ell,k\ell}$ provided*

$$n \geq c \frac{\max\{1, \|\mathbb{E}[T]\|_{op}\}}{\varepsilon \Delta_\ell^2} \cdot \mathbb{E}\left[\max_{i \in [n]} \|a_i\|_2^2\right] \cdot \log k\ell,$$

for an absolute constant c.

The result above applies to any setting where we have i.i.d. samples $a_1, \ldots, a_n$ whose covariance has a block structure under permutation, and the goal is to group the coordinates of a_i based on the unknown block structure. We provide the proof of Theorem 6 at the end of this section.

The next two results provide us with bounds on the spectral quantities $\|\mathbb{E}[T]\|_{op}$ and Δ_ℓ, and on $\mathbb{E}\left[\max_{i\in[n]} \|a_i\|_2^2\right]$ appearing in Theorem 6.

Lemma 7. *Under Assumptions 1 and 2, we have*

$$\|\mathbb{E}[T]\|_{op} \leq \rho \frac{k^2 \ell}{m^2} + \lambda_0^2 \frac{k^3 \ell}{m^2}, \text{ and } \Delta_\ell \geq \frac{\lambda_0^2 k}{\ell}.$$

Lemma 8. *For every $q \in \mathbb{N}$ and $i \in [n]$, we have $\mathbb{E}[\|a_i\|_2^q] \leq c_0^q (\Gamma(q))^2 \lambda_0^q \left(\frac{k\sqrt{k\ell}}{m}\right)^q$. Further, when $\log k \geq 2$, it follows that $\mathbb{E}\left[max_{i\in[n]} \|a_i\|_2^2\right] \leq n^{\frac{2}{\log k}} \mathbb{E}\left[\|a_1\|_2^{\log k}\right]^{\frac{2}{\log k}}$.*

The proof of Lemma 7 is provided in the supplementary material [29] and the proof of Lemma 8 appears in Appendix A-B. We close this section with the proof of Theorem 6.

Proof of Theorem 6. Recall that the proof is conditioned on exact recovery of the union $\mathcal{S}_{\text{un}}$. Further, for notational simplicity, we set $\mathcal{S}_{\text{un}} = [k\ell]$. We divide the proof into two steps. *Step 1. Relating probability of error to perturbation.*

Denote the event that Algorithm 1 labels more than $\varepsilon k\ell$ coordinates incorrectly by $\mathcal{E}$. The following result relates the error probability to a perturbation bound.

Lemma 9 (Error to perturbation bound)**.** *Let V and $\hat{V}$, respectively, be $k\ell \times \ell$ matrices with ith column given by v_i and $\hat{v}_i$, $1 \leq i \leq \ell$, where $v_1, \ldots, v_\ell$ and $\hat{v}_1, \ldots, \hat{v}_\ell$ denote the normalized eigenvectors of $\mathbb{E}[T]$ and T, respectively, corresponding to their ℓ largest eigenvalues. Then,*

$$\Pr(\mathcal{E}) \leq \Pr\left(\|\hat{V} - VO\|_F \geq \frac{1}{2}\sqrt{\frac{\varepsilon\ell}{2}}\right), \tag{7}$$

where $O \in \mathbb{R}^{\ell \times \ell}$ is a random orthonormal matrix and the probability on the right hand side is over the joint distribution of $\hat{V}$ and O.

The proof of this lemma builds on the analysis in [31] and requires us to use some properties of V, which we note in the lemma below.

Lemma 10 (Properties of V)**.** *For $1 \leq i \leq k\ell$, denote by v^i the ith row of V. Then, the following properties hold:*

1) (Identity of rows of V capture the partition) *$v^i = v^j$ if and only if i and j belong to the same support, i.e., $i, j \in \mathcal{S}_t$ for some $t \in [\ell]$.*
2) (Minimum distance property) *For any two distinct rows v^i and v^j, $\|v^i - v^j\|_2^2 \geq 2/k$.*

We provide the proof of Lemma 10 in Appendix A-C.

Proof of Lemma 9. We begin by observing that it suffices to show that

$$\Pr(\mathcal{E}) \leq \Pr\left(\|C - VO\|_F \geq \sqrt{\frac{\varepsilon\ell}{2}}\right), \tag{8}$$

where C is the matrix found in Step 6 of Algorithm 1 and is random since $\hat{V}$ is random. Indeed, by Lemma 10, V has ℓ distinct rows, whereby VO, too, has ℓ distinct rows since O is orthonormal. That is, $VO \in \mathcal{U}_\ell$. Therefore, by triangle inequality, we get

$$\|C - VO\|_F \leq \|C - \hat{V}\|_F + \|VO - \hat{V}\|_F \tag{9}$$

$$= \min_{U \in \mathcal{U}_\ell} \|U - \hat{V}\|_F + \|VO - \hat{V}\|_F \tag{10}$$

$$\leq 2\|VO - \hat{V}\|_F, \tag{11}$$

where the final bound holds since VO belongs to $\mathcal{U}_\ell$. Thus, (8) will imply (7). Note that even if the matrix O were to depend on V and $\hat{V}$ and therefore be random, the result above holds with probability one, and the only property we require from O is orthonormality.

It remains to establish (8). To that end, we define

$$\mathcal{I} \stackrel{\text{def}}{=} \{i \in [k\ell] : \|v^i O - c^i\|_2 < 1/\sqrt{2k}\}, \tag{12}$$

where v^i and c^i are the ith row of V and C, respectively. Our claim is that Algorithm 1 does not make an error in labeling the coordinates in $\mathcal{I}$, unless $|\mathcal{I}^c| > \varepsilon k\ell$. To see this, note that for any two distinct indices $i, j \in \mathcal{I}$ we have

$$\|v^i O - v^j O\|_2 \leq \|v^i O - c^j\|_2 + \|v^j O - c^j\|_2 \tag{13}$$

$$\leq \|v^i O - c^i\|_2 + \|c^i - c^j\|_2 + \|v^j O - c^j\|_2 \tag{14}$$

$$< \sqrt{\frac{2}{k}} + \|c^i - c^j\|_2. \tag{15}$$

Thus, if $c^i = c^j$, we must have $\|v^i O - v^j O\|_2 < \sqrt{2/k}$, which by the second property in Lemma 10 implies that $v^i O =$

v^jO. Therefore, when the labels given by the algorithm for coordinates i and j coincide (this happens only when $c^i = c^j$), then $v^iO = v^jO$. But then, by the first property in Lemma 10, the coordinates i and j must have been in the same part of $\mathcal{S}$.

We have shown that the indices in $\mathcal{I}$ that are assigned the same label by the algorithm must come from the same part in $\mathcal{S}$. We still need to verify that coordinates from the same part in $\mathcal{S}$ do not get assigned to different parts. We show this cannot happen unless $|\mathcal{I}^c| > \varepsilon k\ell$, and this is where we use the assumption that $\varepsilon < 1/\ell$. Indeed, if $|\mathcal{I}^c| \le \varepsilon k\ell < k$, then at least one element from each part $\mathcal{S}_1, \dots, \mathcal{S}_\ell$ must be in $\mathcal{I}$, since $|\mathcal{S}_i| = k$ for every i. By our previous observation, elements in each of these parts in $\mathcal{I}$ must be assigned different labels by the algorithm, which means that it must assign at least ℓ different labels to the elements in $\mathcal{I}$. Thus, if the algorithm assigns two elements in the same part $\mathcal{S}_i$ different labels, it will assign more that ℓ different labels, which is not allowed.

Therefore, all the indices in $\mathcal{I}$ are correctly labeled when $|\mathcal{I}^c| \le \varepsilon k\ell$. Then, clearly, in this case the error event $\mathcal{E}$ does not hold. It follows from the definition of $\mathcal{I}$ that

$$\Pr(\mathcal{E}) \le \Pr\left(|\mathcal{I}^c| > \varepsilon k\ell\right) \tag{16}$$

$$\le \Pr\left(\left|\left\{i : \|c^i - v^iO\|_2 \ge \frac{1}{\sqrt{2k}}\right\}\right| > \varepsilon k\ell\right) \tag{17}$$

$$\le \Pr\left(\|C - VO\|_F^2 > \frac{\varepsilon\ell}{2}\right), \tag{18}$$

where in the final step we used the fact that the second step implies $\|C - VO\|_F^2 = \sum_{i=1}^{k\ell} \|c^i - v^iO\|_2^2 \ge \varepsilon k\ell/2k$. This completes the proof of (8). □

Step 2: Controlling the perturbation.

In view of Lemma 9, we only need to control the perturbation $\|\hat{V} - VO\|_F$. We do this using the following extension of the Davis-Kahan theorem, which also fixes the choice of O.

Theorem 11 (Perturbation of eigenspace)**.** *[36] Let A and $\hat{A}$ be $d \times d$ symmetric matrices with eigenvalues $\nu_1 \ge \cdots \ge \nu_d$ and $\hat{\nu}_1 \ge \cdots \ge \hat{\nu}_d$, respectively. Let V and $\hat{V}$ be $d \times \ell$ matrices consisting of the ℓ leading normalized eigenvectors of A and $\hat{A}$, respectively. Then, there exists an orthonormal matrix $O \in \mathbb{R}^{\ell\times\ell}$ such that*

$$\|\hat{V} - VO\|_F^2 \le 2\sqrt{2}\,\frac{\min\{\sqrt{\ell}\|\hat{A} - A\|_{op}, \|\hat{A} - A\|_F\}}{\nu_\ell - \nu_{\ell+1}}. \tag{19}$$

By applying this result with T and $\mathbb{E}[T]$ in the role of $\hat{A}$ and A, respectively, we get that there exists an orthonormal matrix O such that

$$\|\hat{V} - VO\|_F \le \frac{2\sqrt{2}}{\Delta_\ell}\min\{\sqrt{\ell}\|T - \mathbb{E}[T]\|_{op}, \|T - \mathbb{E}[T]\|_F\}, \tag{20}$$

where $\Delta_\ell \stackrel{\text{def}}{=} \nu_\ell - \nu_{\ell+1}$. Combining this bound with our earlier bound from Lemma 9, we get

$$\Pr(\mathcal{E}) \le \Pr\left(\|T - \mathbb{E}[T]\|_{op} \ge \frac{\Delta_\ell\sqrt{\varepsilon}}{8}\right) \tag{21}$$

$$\le \frac{8}{\Delta_\ell\sqrt{\varepsilon}} \cdot \mathbb{E}\left[\|T - \mathbb{E}[T]\|_{op}\right], \tag{22}$$

where the last step uses Markov's inequality.

To bound the expected value on the right hand side, we use the following extension of a result of Rudelson [34]. As pointed out earlier, the original bound in [34] was restricted to isotropic Z_is, and we show that it extends to arbitrary i.i.d. Z_is with an extra factor. The proof is provided in Appendix A-D.

Theorem 12 (Extension of a result in [34])**.** *Let $Z \in \mathbb{R}^N$ be a random vector such that $A = \mathbb{E}\left[ZZ^\top\right]$. Let $Z_1, \dots, Z_n$ be independent copies of Z. Then, there exists an absolute constant c such that*

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_iZ_i^\top - A\right\|_{op}\right] \le \frac{1}{2}\left(\alpha^2 + \alpha\sqrt{\alpha^2 + 4\|A\|_{op}}\right), \tag{23}$$

where

$$\alpha = c\sqrt{\frac{\mathbb{E}\left[\max_{i\in[n]}\|Z_i\|_2^2\right]\log N}{n}}.$$

Using this bound in (22) with $N = k\ell$, we obtain

$$\Pr(\mathcal{E}) \le \frac{4}{\Delta_\ell\sqrt{\varepsilon}}\left(\alpha^2 + \alpha\sqrt{\alpha^2 + 4\|\mathbb{E}[T]\|_{op}}\right). \tag{24}$$

The proof is completed upon noting that α can be made smaller than $1/2$ using $n \ge c\mathbb{E}\left[\max_{i\in[n]}\|a_i\|_2^2\right]\log k\ell$, in which case $\alpha\sqrt{\alpha^2 + 4\|\mathbb{E}[T]\|_{op}} \le \alpha\sqrt{8\max\{1, \|\mathbb{E}[T]\|_{op}\}}$. The error probability above can thus be made less than $1/4$ if $n \ge c(\log k\ell)\max\{1, \|\mathbb{E}[T]\|_{op}\}\mathbb{E}\left[\max_{i\in[n]}\|a_i\|_2^2\right]/(\Delta_\ell^2\varepsilon)$. □

In the next section, we combine the results from Theorems 3 and 6 to show the sample complexity bound of Theorem 1.

C. Proof of Theorem 1

The proof of Theorem 1 now follows by combining guarantees for the union recovery step from Theorem 3 and the clustering step from Theorem 6.

We begin by applying Theorem 3 to get that $\hat{\mathcal{S}}_{\text{un}}$ coincides with $\mathcal{S}_{\text{un}} = \cup_{i=1}^\ell \mathcal{S}_i$ with probability close to 1. Throughout, we condition on this event occurring. However, to avoid technical difficulties, we assume that a different set of independent samples is used to recover $\mathcal{S}_{\text{un}}$ than those used to recover $\mathcal{S}_1, \dots, \mathcal{S}_\ell$ – thus, the overall number of samples needed will be the sum of samples needed for union recovery in Theorem 3 and the sample complexity determined in our analysis below. In particular, the clustering step dominates the sample complexity of our algorithm.

Next, upon substituting the bounds from Lemma 7 and Lemma 8 into Theorem 6, we see that for ε-approximate recovery of the supports it suffices to have

$$n \ge \frac{c}{\varepsilon}\lambda_0^2\frac{k^3\ell}{m^2}\frac{\ell^2}{\lambda_0^4k^2} \cdot n^{\frac{2}{\log k}} \times \left(\lambda_0\frac{k\sqrt{k}\sqrt{\ell}}{m}(\log k)^2\right)^2 \cdot \log(k\ell) \tag{25}$$

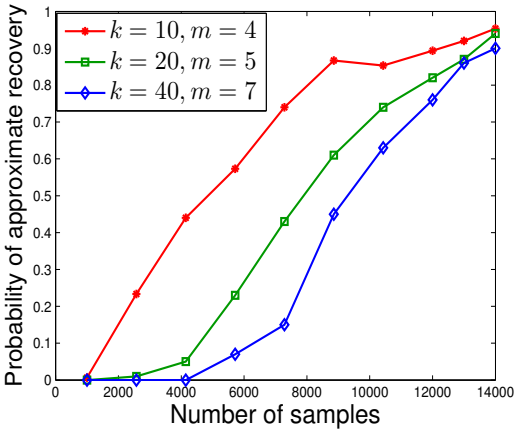


Fig. 2: Probability of approximate support recovery with $d = 100$, $\varepsilon = 0.2$, $\ell = 2$, and varying k/m ratios.

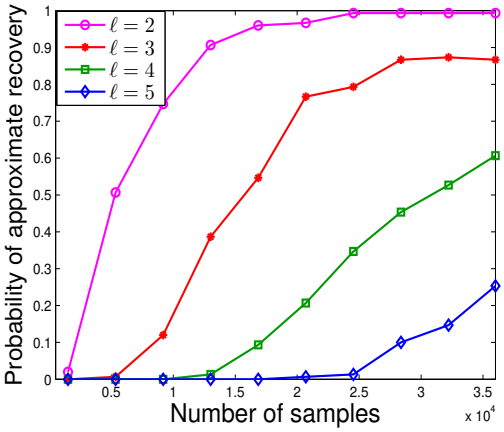


Fig. 3: Probability of approximate support recovery with $d = 100$, $\varepsilon = 0.2$, $m = 4$, $k = 10$, and varying ℓ.

$$= \frac{c}{\varepsilon}\frac{k^4\ell^4}{m^4} n^{\frac{2}{\log k}} (\log k)^4 \log(k\ell). \tag{26}$$

For $n \geq c((1/\varepsilon)(k\ell/m)^4 \cdot (\log k)^4 \log(k\ell))$, $n^{\frac{1}{\log k}} = O(1)$, which completes the proof in view of the sufficient condition for n above.

V. Simulations

A. Synthetic data

In this subsection, we evaluate the performance of Algorithm 1 on synthetic data for various parameter values. Through these simulations, our goal is to see how the performance of the algorithm varies as a function of the ratio k/m and ℓ for a fixed d.

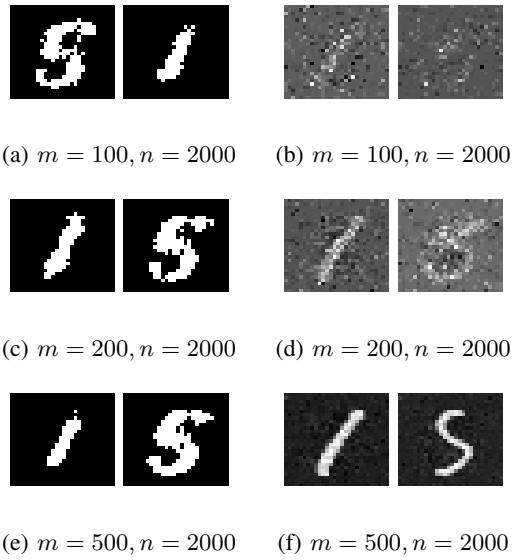
We first choose $d = 100$, $\ell = 2$ and consider three different values of k/m. We generate two disjoint subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of $[d]$, each of size k. Then, for a given n, we generate $n/2$ samples with each support, with values on the support drawn from the standard normal distribution in $\mathbb{R}^k$. Measurement matrices $\{\Phi_i\}_{i=1}^n$ are generated independently with i.i.d. $\mathcal{N}(0, 1/m)$ entries and multiplied with the samples to obtain measurements $\{Y_i\}_{i=1}^n$. These measurements are given as input to the support recovery algorithm, which produces estimates for the union, as well as the individual supports, which we denote by $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$. For each value of (k, m, n), we run 100 trials and declare it a success if the error $\sum_{i=1}^2 |\hat{\mathcal{S}}_i \Delta \mathcal{S}_{\sigma(i)}| < 2\varepsilon k$. The plot in Figure 2 shows the success rate over the 100 trials as a function of the number of samples n, with ε set as 0.2. Note that the number of measurements taken per sample, m, is much smaller than the support size, k, of each sample. We can see from Figure 2 that for a fixed probability of success, the number of samples required increases with k/m, which agrees with the result in Theorem 1. In Figure 3, we show the variation in the probability of approximate recovery as a function of n for the number of supports $\ell = \{2, 3, 4, 5\}$, with k and m (and hence their ratio) held fixed. We can see that the number of samples required to achieve a given probability of recovery increases with ℓ. Our current experiments however do not reveal whether the dependence on these parameters is tight.

B. MNIST dataset

As an application involving natural data, we consider the problem of reconstructing handwritten images from very few linear measurements. We apply the multiple support recovery algorithm to the MNIST dataset [37], which consists of $60,000$ images of handwritten digits, each of size 28×28. Each (grayscale) image is a sample in our setting, and the support of the sample essentially identifies the digit. This dataset fits well into our hypothesis that there is a small set of unknown supports underlying the data – handwritten images corresponding to the same digit can be thought of as having roughly the same pattern (support) in the pixel domain. Thus, the vectorized version of images of the same digit will have approximately the same support. We note that the task here is to recover the images of the digits from low dimensional projections, and not to learn a classifier using the dataset.

In our experiments, the vectorized version of each image (a 784×1 vector) is projected onto $m = 100, 200$ or 500 dimensions using Gaussian measurement matrices described in Assumption 2. Given these low dimensional projections, the goal is to identify the underlying digits. We fix $\ell = 2$ and consider the example of digits 1 and 5 as shown in Figure 4. The support size of each digit is roughly in the range $150 - 200$. It can be seen that Algorithm 1 can identify the distinct digits even when $m < k$. For comparison, we used the Group LASSO algorithm on the projected samples, which tries to recover the individual samples (images) itself. However, it requires a much larger number of measurements per sample (for example, about $m = 500$ in this case). In fact, previously known algorithms for sparse recovery do not perform well in the low measurement regime of $m < k$, and we have used Group LASSO as an example to illustrate this fact.

We note that since these are handwritten digits, the support of samples coming from the same digit can also vary to some extent. However, the averaging across samples in our estimator takes care of this problem. Further, the supports from different digits need not be disjoint. To handle overlaps, we



(a) $m = 100, n = 2000$ (b) $m = 100, n = 2000$

(c) $m = 200, n = 2000$ (d) $m = 200, n = 2000$

(e) $m = 500, n = 2000$ (f) $m = 500, n = 2000$

Fig. 4: Recovery performance of Algorithm 1 ((a),(c),(e)), and Group LASSO ((b),(d),(f)).

use the observation that $\tilde{\lambda}$ can provide an estimate for the intersection of supports as well. The plot of sorted entries of $\tilde{\lambda}$ shows a sharp drop in values at two locations, one around the intersection and another around the union. We include this estimate of intersection of supports into our final estimate. This method performs well in practice, as can be seen in the results of Figure 4, where digits 1 and 5 have significant overlap.

C. Computational complexity

The first step in our algorithm for estimating the union involves computing the average variance along each of the d coordinates and requires $O(mnd)$ operations. The clustering step involves computing the T matrix and its ℓ leading eigenvectors which requires $O(k^3\ell^3 + k^2\ell^2 n)$ operations, followed by the ℓ-means step which requires $O(k\ell^3)$ operations per iteration. Other algorithms for recovering multiple supports do not perform well when $m < k$, and have computational complexity that scales quadratically or worse with d. For instance, the sparse Bayesian learning based algorithm from [17] has a complexity of $O(d^2)$ per iteration, and LASSO-based procedures have a complexity of $O(d^2)$ or $O(d^3)$ per iteration, depending on the specific algorithm used.

VI. Discussion

Throughout in this work, we assumed that the distinct supports were pairwise disjoint sets. In the case of overlapping supports, the structure of the expected affinity matrix, and consequently its spectrum, changes. For the special case of $\ell = 2$, overlapping supports can be handled by a simple modification of the sign-based estimate. Instead of partitioning the coordinates in the union estimate based on the sign of the eigenvector, we now use a threshold $\tau > 0$ and declare coordinates with values in $[-\tau, \tau]$ as belonging to both supports (values above τ or below $-\tau$ are assigned to different supports). The optimal τ can be explicitly characterized in terms of the parameters of the problem. Given our current algorithm, a simple way to handle this case for *general* ℓ would be to use fuzzy ℓ-means, which returns scores for each coordinate indicating how likely it is to belong to a certain support. However, choosing a threshold to decide the supports using the scores is difficult in general. Some other approaches have been explored in the graph clustering literature, but these do not apply directly to our setting. Other extensions of this work include studying the performance of the algorithm under different support sizes, and prior distribution with non-uniform mixing weights. Also, our work shows a sufficient condition on the number of samples required for multiple support recovery; obtaining the necessary condition is a challenging task in general and requires characterizing the distance between mixture distributions. Using a component wise distance bound leads to the same lower bound as in [4] (with an additional $1/\ell$ factor), and obtaining a better lower bound seems difficult.

Appendix A
Remaining proofs from Section IV-B

A. Proof of Lemma 5 (Probability of error boosting)

Given an $(n, \varepsilon, 1/4)$-estimator for $\Sigma_{k,\ell,d}$, we apply it to L independent blocks of data. Specifically, denoting this estimator by e, consider independent copies $(Y^n(t), \Phi^n(t))$, $1 \le t \le L$, of (Y^n, Φ^n). For $t \in [L]$, let

$$(\hat{\mathcal{S}}_{1,t}, \ldots, \hat{\mathcal{S}}_{\ell,t}) := e(Y^n(t), \Phi^n(t))$$

denote the output for the estimator applied to the tth block.

We now describe a procedure to output a final estimate for the supports using the estimates $(\hat{\mathcal{S}}_{1,t}, \ldots, \hat{\mathcal{S}}_{\ell,t})$ from the L blocks of samples. For each $t \in [L]$, we check if there is a set $\mathcal{I} \subseteq [L]\backslash\{t\}$ of cardinality $N \ge L/2$ satisfying

$$\min_{\sigma_t \in \mathcal{G}_\ell} \frac{1}{k\ell} \sum_{i=1}^{\ell} |\hat{\mathcal{S}}_{i,t} \Delta \hat{\mathcal{S}}_{\sigma_t(i),t'}| \le 2\varepsilon, \quad \forall\, t' \in \mathcal{I}. \tag{27}$$

That is, we look for a t for which $(\hat{\mathcal{S}}_{1,t}, \ldots, \hat{\mathcal{S}}_{\ell,t})$ are close to $L/2$ other estimates. This indicates "robustness" of the estimate from the tth block, making it an appropriate proxy for the median. Our final estimate is $(\bar{\mathcal{S}}_1, \ldots, \bar{\mathcal{S}}_\ell) = (\hat{\mathcal{S}}_{1,t}, \ldots, \hat{\mathcal{S}}_{\ell,t})$, where t is an index which satisfies the property above.

We show that for $L \ge \lceil 8 \ln \frac{1}{\delta} \rceil$ the estimator above constitutes an $(nL, 3\varepsilon, \delta)$-estimator for $\Sigma_{k,\ell,d}$. Indeed, denoting

$$Z_t = \mathbb{1}\left(\exists\, \sigma \in \mathcal{G}_\ell \text{ s.t. } \frac{1}{k\ell} \sum_{i=1}^{\ell} |\mathcal{S}_i \Delta \hat{\mathcal{S}}_{\sigma(i),t}| \le \varepsilon \right),$$

by our assumption for the estimator e we have

$$\mathbb{E}_{\mathrm{P}_{(\mathcal{S}_1,\ldots,\mathcal{S}_\ell)}}[Z_t] \ge \frac{3}{4}.$$

Furthermore, Z_t are independent for different $t \in [L]$. Thus, by Hoeffding's inequality,

$$\mathrm{P}_{(\mathcal{S}_1,\ldots,\mathcal{S}_\ell)}\left(\sum_{t=1}^{L} Z_t \le \frac{L}{2} \right) \le e^{-\frac{L}{8}}, \quad \forall\, (\mathcal{S}_1, \ldots, \mathcal{S}_\ell) \in \Sigma_{k,\ell,d}.$$

In particular, for $L \geq \lceil 8 \ln \frac{1}{\delta} \rceil$, with probability exceeding $1-\delta$ there exist[5] $M \geq L/2+1$ indices $t_1, \ldots, t_M \in [L]$ and permutations $\sigma_1, \ldots, \sigma_M \in \mathcal{G}_\ell$ such that

$$\frac{1}{k\ell} \sum_{i=1}^{\ell} |\mathcal{S}_i \Delta \hat{\mathcal{S}}_{\sigma_j(i),t_j}| \leq \varepsilon, \quad \forall\, j \in [M]. \tag{28}$$

Note that since $|A \Delta B|$ is a metric for subsets of $[d]$, the estimate $(\hat{\mathcal{S}}_{1,t}, \ldots, \hat{\mathcal{S}}_{\ell,t})$ for $t = t_1$ satisfies (27) when (28) holds; in fact, any index among $\{t_1, ..., t_M\}$ can serve this purpose. However, the estimate described earlier need not select any of these indices. Yet, we now show that any other index chosen by the procedure will work as well, provided (28) holds.

To that end, denote by $\mathcal{I}'$ the set $\{t_1, \ldots, t_M\}$ of indices satisfying (28), and recall the set $\mathcal{I}$ found by our estimation procedure earlier. Then, when $|\mathcal{I}'| \geq L/2+1$, which holds with probability exceeding $1-\delta$,

$$|\mathcal{I} \cap \mathcal{I}'| \geq |\mathcal{I}| + |\mathcal{I}'| - L \geq 1,$$

whereby there exists an index $t \in [L]$ and permutations $\sigma, \overline{\sigma} \in \mathcal{G}_\ell$ such that

$$\frac{1}{k\ell} \sum_{i=1}^{\ell} |\mathcal{S}_i \Delta \hat{\mathcal{S}}_{\sigma(i),t}| \leq \varepsilon \quad \text{and} \quad \frac{1}{k\ell} \sum_{i=1}^{\ell} |\overline{\mathcal{S}}_i \Delta \hat{\mathcal{S}}_{\overline{\sigma}(i),t}| \leq 2\varepsilon.$$

It follows that the permutation $\sigma' = \sigma \circ \overline{\sigma}^{-1}$ satisfies

$$\frac{1}{k\ell} \sum_{i=1}^{\ell} |\mathcal{S}_i \Delta \overline{\mathcal{S}}_{\sigma'(i)}| \leq 3\varepsilon,$$

which completes the proof. □

B. *Proof of Lemma 8*

As noted in the proof of Theorem 1, the clustering step in our algorithm is analyzed under the assumption that the union of supports is exactly recovered in the first step, whereby we can set $\hat{\mathcal{S}}_{\text{un}} = \mathcal{S}_{\text{un}}$.

We will first show the bound on $\mathbb{E}\left[\max_{i\in[n]} \|a_i\|_2^2\right]$, followed by the moment bound for $\mathbb{E}\left[\|a_i\|_2^q\right]$. We start by noting that for any $q \geq 2$,

$$\mathbb{E}\left[\max_{i\in[n]} \|a_i\|_2^2\right] = \mathbb{E}\left[\left(\max_{i\in[n]} \|a_i\|_2^q\right)^{\frac{2}{q}}\right] \tag{29}$$

$$\leq \mathbb{E}\left[\left(\sum_{i=1}^{n} \|a_i\|_2^q\right)^{\frac{2}{q}}\right] \tag{30}$$

$$\leq \left(\mathbb{E}\left[\sum_{i=1}^{n} \|a_i\|_2^q\right]\right)^{\frac{2}{q}} \tag{31}$$

$$= n^{\frac{2}{q}} \left(\mathbb{E}\left[\|a_1\|_2^q\right]\right)^{\frac{2}{q}}, \tag{32}$$

where we used Jensen's inequality in the third step. For $\log k \geq 2$, upon setting $q = \log k$ in the inequality above, we get

$$\mathbb{E}\left[\max_{i\in[n]} \|a_i\|_2^2\right] \leq n^{\frac{2}{\log k}} \left(\mathbb{E}\left[\|a_1\|_2^{\log k}\right]\right)^{\frac{2}{\log k}}. \tag{33}$$

We now proceed to bound $\mathbb{E}\left[\|a_i\|_2^q\right]$. In the rest of the proof, we will denote $a_i \in \mathbb{R}^d$ by a, and with some abuse of notation, denote by Φ_i the ith column of Φ. By using the definition of a, we have

$$\|a\|_2^{2q} = \left(\sum_{i\in\mathcal{S}_{\text{un}}} a_i^2\right)^q \tag{34}$$

$$= \left(\sum_{i\in\mathcal{S}_{\text{un}}} (\Phi_i^\top \Phi_{\mathcal{S}} X_{\mathcal{S}})^4\right)^q \tag{35}$$

$$= \left(\sum_{i\in\mathcal{S}_{\text{un}}} (\alpha_i^\top X_{\mathcal{S}})^4\right)^q \tag{36}$$

$$= \left(\sum_{i\in\mathcal{S}_{\text{un}}} (X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\right)^q, \tag{37}$$

where $\alpha_i = \Phi_{\mathcal{S}}^\top \Phi_i$ as defined before and $A_i \stackrel{\text{def}}{=} \alpha_i \alpha_i^\top$. To compute the expectation of the term in the last step, we first condition on Φ and note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i\in\mathcal{S}_{\text{un}}} (X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\right)^q \middle| \Phi\right] \\
&= (k\ell)^q \mathbb{E}\left[\left(\frac{1}{k\ell} \sum_{i\in\mathcal{S}_{\text{un}}} (X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\right)^q \middle| \Phi\right] \\
&\leq (k\ell)^{q-1} \sum_{i\in\mathcal{S}_{\text{un}}} \mathbb{E}\left[(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^{2q} | \Phi\right],
\end{aligned} \tag{38}$$

where we used $|\mathcal{S}_{\text{un}}| = k\ell$, and the convexity of the function x^q for $x \geq 0$, $q \in \mathbb{N}$. The quantity on the right essentially involves the $(2q)$th moment of a subexponential random variable (see Appendix B for definition). To see that the quadratic form $X_{\mathcal{S}}^\top A_i X_{\mathcal{S}}$ is subexponential, we use the Hanson-Wright inequality (*cf.* [38]) to get

$$\begin{aligned}
&\mathbb{P}(|X_{\mathcal{S}}^\top A_i X_{\mathcal{S}} - \mu| \geq t | \Phi) \\
&\leq 2 \exp\left(-\min\left\{\frac{t^2}{\lambda_0^2 \|A_i\|_F^2}, \frac{t}{\lambda_0 \|A_i\|_{op}}\right\}\right),
\end{aligned} \tag{39}$$

where $\mu = \mathbb{E}\left[X_{\mathcal{S}}^\top A_i X_{\mathcal{S}} | \Phi\right] = \lambda_0 \|\alpha_i\|_2^2$. Lemma 13 in Appendix B can now be used to bound the moment in (38). Specifically, we get

$$\begin{aligned}
&\mathbb{E}[(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^{2q} | \Phi] \\
&\leq 2q \cdot (16)^q \left(\Gamma(q) \lambda_0^{2q} \|A_i\|_F^{2q} + \Gamma(2q) \lambda_0^{2q} \|A_i\|_{op}^{2q}\right) + 2^{2q} \mu^{2q}
\end{aligned} \tag{40}$$

$$\leq 3q \cdot (16)^q \Gamma(2q) \lambda_0^{2q} \|\alpha_i\|_2^{4q}, \tag{41}$$

where we used $\|A_i\|_F = \|A_i\|_{op} = \|\alpha_i\|_2^2$. Next, taking expectation over Φ, we obtain

$$\mathbb{E}\left[(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^{2q}\right] \leq c_q' \Gamma(2q) \lambda_0^{2q} \mathbb{E}\left[\|\alpha_i\|_2^{4q}\right], \tag{42}$$

where $c_q' = 3q \cdot (16)^q$. Thus, combining the result above with

[5] Without loss of generality, we assume L to be even.

(38), we get

$$\mathbb{E}\left[\left(\sum_{i\in\mathcal{S}_{\text{un}}}(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\right)^q\right] \le c_q'\Gamma(2q)\lambda_0^{2q}(k\ell)^q\sum_{i\in\mathcal{S}_{\text{un}}}\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right] \tag{43}$$

$$= c_q'\Gamma(2q)\lambda_0^{2q}(k\ell)^q\left(\sum_{i\in\mathcal{S}}\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right] + \sum_{i\in\mathcal{S}_{\text{un}}\setminus\mathcal{S}}\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right]\right). \tag{44}$$

When $i\in\mathcal{S}$,

$$\mathbb{E}[\|\alpha_i\|_2^{4q}] = \mathbb{E}\left[\left(\|\Phi_i\|_2^4 + \sum_{j\in\mathcal{S}\setminus\{i\}}(\Phi_i^\top\Phi_j)^2\right)^{2q}\right] \tag{45}$$

$$\le 2^{2q}\left(\mathbb{E}\left[\|\Phi_i\|_2^{8q}\right] + \mathbb{E}\left[\left(\sum_{j\in\mathcal{S}\setminus\{i\}}(\Phi_i^\top\Phi_j)^2\right)^{2q}\right]\right), \tag{46}$$

and when $i\in\mathcal{S}_{\text{un}}\setminus\mathcal{S}$,

$$\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right] \le \mathbb{E}\left[\left(\sum_{j\in\mathcal{S}}(\Phi_i^\top\Phi_j)^2\right)^{2q}\right]. \tag{47}$$

Since Φ_i has independent, subgaussian entries with parameter $1/m$, we see that $\|\Phi_i\|_2^2 \sim \text{subexp}(c'/m, c''/m)$ with $c'=128$ and $c''=8$ [4, Lemma D.2]. This gives, using Lemma 13,

$$\mathbb{E}\left[(\|\Phi_i\|_2^2)^{4q}\right] \le 2q(16)^q\left(\Gamma(2q)\frac{c'^{2q}}{m^{2q}} + \Gamma(4q)\frac{c''^{4q}}{m^{4q}}\right) + (\mathbb{E}\left[\|\Phi_i\|_2^2\right])^{4q} \tag{48}$$

$$\le 4q(16)^q c'^{2q}\Gamma(4q)\frac{1}{m^{2q}} + 1, \tag{49}$$

where we used $c' > c''^2$. Using similar arguments, we note that $\Phi_i^\top\Phi_j|\Phi_i$ is subgaussian with parameter $\|\Phi_i\|_2^2/m$, which implies that, conditioned on Φ_i, $\sum_{j\in\mathcal{S}\setminus\{i\}}(\Phi_i^\top\Phi_j)^2$ is subexp$(c'(k-1)\|\Phi_i\|_2^4/m^2, c''\|\Phi_i\|_2^2/m)$. Then, using Lemma 13 again, we get

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{j\in\mathcal{S}\setminus\{i\}}(\Phi_i^\top\Phi_j)^2\right)^{2q}\right] \\
&\le c_q'\mathbb{E}_{\Phi_i}\left[\Gamma(q)c'^q\left(\frac{k-1}{m^2}\right)^q\|\Phi_i\|_2^{4q} + \Gamma(2q)c''^{2q}\left(\frac{\|\Phi_i\|_2^2}{m}\right)^{2q}\right] \\
&\quad + 2^{2q}\left(\mathbb{E}\left[\sum_{j\in\mathcal{S}\setminus\{i\}}(\Phi_i^\top\Phi_j)^2\right]\right)^{2q} \\
&\le c_q'c'^q\Gamma(q)\left(\frac{k-1}{m^2}\right)^q\left(1 + 2c_q'c'^{2q}\Gamma(2q)\frac{1}{m^q}\right) \\
&\quad + c_q'c''^{2q}\Gamma(2q)\frac{1}{m^{2q}}\left(1 + c_q'c'^{2q}\Gamma(2q)\frac{1}{m^q}\right) + 2^{2q}\left(\frac{k-1}{m}\right)^{2q} \\
&\le 5c_q'c'^{2q}\Gamma(2q)\left(\frac{k}{m}\right)^{2q}.
\end{aligned}$$

Combining these results and substituting into (44), we get

$$\mathbb{E}\left[\left(\sum_{i\in\mathcal{S}_{\text{un}}}(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\right)^q\right] \le c_q'\Gamma(2q)\lambda_0^{2q}(k\ell)^{q-1}\left(\sum_{i\in S}\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right] + \sum_{i\in\mathcal{S}_{\text{un}}\setminus\mathcal{S}}\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right]\right) \tag{50}$$

$$\le 5c_q'^2c'^{2q}\Gamma(2q)\lambda_0^{2q}(k\ell)^{q-1}\left(k\Gamma(2q)\left(\frac{k}{m}\right)^{2q} + (k\ell-k)\Gamma(2q)\left(\frac{k}{m}\right)^{2q}\right) \tag{51}$$

$$= 5c_q'^2c'^{2q}(\Gamma(2q))^2\lambda_0^{2q}\left(\frac{k\sqrt{k\ell}}{m}\right)^{2q}. \tag{52}$$

Rescaling the exponent, we get

$$\mathbb{E}\left[\|a\|_2^q\right] = \mathbb{E}\left[\left(\sum_{i\in\mathcal{S}_{\text{un}}}(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\right)^{\frac{q}{2}}\right] \tag{53}$$

$$\le 5c_{q/2}^2c'^q(\Gamma(q))^2\lambda_0^q\left(\frac{k\sqrt{k\ell}}{m}\right)^q \tag{54}$$

Noting that $c'(5c_{q/2}^2)^{1/q} \le 45\cdot 8c' = c_0$, we obtain the result. □

C. Proof of Lemma 10

(i) To show the first property, we note that the true covariance matrix can be decomposed as $\mathbb{E}[T] = WBW^\top + (\mu_0-\mu_s)I$, where $W\in\{0,1\}^{k\ell\times\ell}$ encodes the block structure, and $B\in\mathbb{R}^{\ell\times\ell}$ contains the distinct values from each block. In particular, for $1\le i\le k\ell$ and $1\le j\le\ell$, define

$$W_{ij} = \begin{cases}1, & \text{if } i\in\mathcal{S}_j,\\ 0, & \text{otherwise,}\end{cases} \tag{55}$$

and, for $1\le i\le\ell$ and $1\le j\le\ell$, define

$$B_{ij} = \begin{cases}\mu_s, & \text{if } i=j,\\ \mu_d, & \text{otherwise.}\end{cases} \tag{56}$$

Since $\mathbb{E}[T]$ and $WBW^\top$ have the same set of eigenvectors, we will show that the matrix $V\in\mathbb{R}^{k\ell\times\ell}$ consisting of the ℓ leading eigenvectors of $WBW^\top$ has the desired property. To that end, first note that there are only ℓ unique rows in W, one unique row corresponding to each block. We will show that V also consists of ℓ unique rows, in exact correspondence with the rows of W. To do so, we will follow [31, Lemma 3.1] and show that V is essentially a row-transformed version of W, i.e., there exists an invertible matrix $H\in\mathbb{R}^{\ell\times\ell}$ such that $WH=V$. We start by considering the eigen decomposition

$$(W^\top W)^{\frac{1}{2}}B(W^\top W)^{\frac{1}{2}} = U\Lambda U, \tag{57}$$

where $\Lambda\in\mathbb{R}^{\ell\times\ell}$ is diagonal and $U\in\mathbb{R}^{\ell\times\ell}$ is an orthonormal matrix. Left multiplying by $W(W^\top W)^{-\frac{1}{2}}$

and right multiplying by $(W^\top W)^{-\frac{1}{2}}W^\top$ in the equation above, we get,

$$WBW^\top = WH\Lambda(WH)^\top, \tag{58}$$

where $H \stackrel{\text{def}}{=} (W^\top W)^{-\frac{1}{2}}U$. Finally, right multiplying by WH and noting that $(WH)^\top WH = I$, we have

$$WBW^\top \cdot WH = WH \cdot \Lambda, \tag{59}$$

implying that the columns of WH are the normalized eigenvectors of $WBW^\top$.

We have thus shown that $V = WH$. Let v^i and w^i denote the ith row of V and W, respectively. If $v^i = v^j$ for some $i \neq j$, then $w^iH = w^jH$. Since $H = (W^\top W)^{-\frac{1}{2}}U$ is invertible, this implies $w^i = w^j$. Conversely, if $w^i = w^j$ for some $i \neq j$, then $w^iH = w^jH$, which implies $v^i = v^j$.

(ii) Using the fact that $V = WH$ from (i), we have for $v^i \neq v^j$,

$$\|v^i - v^j\|_2 = \|(w^i - w^j)H\|_2 \tag{60}$$
$$\geq \sqrt{2}\nu_{\min}(H), \tag{61}$$

where $\nu_{\min}(H) \stackrel{\text{def}}{=} \min_{\|x\|_2=1}\|x^\top H\|_2$, and we used $\|w^i - w^j\|_2 = \sqrt{2}$ for $w^i \neq w^j$. Now,

$$\min_{\|x\|_2=1}\|x^\top H\|_2^2 = \min_{\|x\|_2=1} x^\top HH^\top x \tag{62}$$
$$= \min_{\|x\|_2=1} x^\top (WW^\top)^{-1}x \tag{63}$$
$$= \frac{1}{k}, \tag{64}$$

where we used $HH^\top = (W^\top W)^{-\frac{1}{2}}UU^\top(WW^\top)^{-\frac{1}{2}} = (WW^\top)^{-1}$ and the fact that $WW^\top = k\,\texttt{diag}\,(I)$. Putting everything together, we get

$$\|v^i - v^j\|_2^2 \geq \frac{2}{k}. \tag{65}$$

D. Proof of Theorem 12

The proof is similar to that of [34], and we highlight the steps needed to extend the result to general A. In particular, following similar arguments as in [34], it can be shown that

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_iZ_i^\top - A\right\|_{op}\right] \leq c\frac{\sqrt{\log N}}{n}\sqrt{\mathbb{E}\left[\max_{i\in[n]}\|Z_i\|_2^2\right]}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^n Z_iZ_i^\top\right\|_{op}\right]}, \tag{66}$$

Now,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Z_iZ_i^\top\right\|_{op}\right] \leq n\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_iZ_i^\top - A\right\|_{op} + \|A\|_{op}\right] \tag{67}$$
$$= n(\beta + \|A\|_{op}), \tag{68}$$

where $\beta \stackrel{\text{def}}{=} \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n Z_iZ_i^\top - A\right\|_{op}\right]$. It follows from (66) and (68) that

$$\beta \leq c\sqrt{\frac{\log N}{n}}\sqrt{\mathbb{E}\left[\max_{i\in[n]}\|Z_i\|_2^2\right]}\sqrt{\beta + \|A\|_{op}}. \tag{69}$$

Letting $\alpha = c\sqrt{(\log N)/n}\sqrt{\mathbb{E}\left[\max_{i\in[n]}\|Z_i\|_2^2\right]}$, we have the solution

$$\beta \leq \frac{1}{2}\left(\alpha^2 + \alpha\sqrt{\alpha^2 + 4\|A\|_{op}}\right), \tag{70}$$

which completes the proof.

Appendix B
Moment and Concentration Bounds For Subgaussian Random Variables

Definition 2. A random variable X is subgaussian with variance parameter σ^2, denoted $X \sim \text{subG}(\sigma^2)$, if

$$\log \mathbb{E}\left[e^{\theta(X-\mathbb{E}[X])}\right] \leq \theta^2\sigma^2/2, \tag{71}$$

for all $\theta \in \mathbb{R}$.

Definition 3. A random variable X is subexponential with parameters σ^2 and $b > 0$, denoted $X \sim \text{subexp}(\sigma^2, b)$, if

$$\log \mathbb{E}\left[e^{\theta(X-\mathbb{E}[X])}\right] \leq \theta^2\sigma^2/2, \tag{72}$$

for all $|\theta| < 1/b$.

Lemma 13. *Let X be a subexponential random variable with parameters v^2 and $b > 0$, i.e., for every $t > 0$,*

$$\Pr\left(|X - \mathbb{E}[X]| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2}{2v^2}, \frac{t}{2b}\right\}\right). \tag{73}$$

Then, for $q \in \mathbb{N}$, and an absolute constant c,

$$\mathbb{E}\left[|X - \mathbb{E}[X]|^{2q}\right] \leq 2q\cdot(16)^q\left(\Gamma(q)v^{2q} + b^{2q}\Gamma(2q)\right). \tag{74}$$

Proof. We first express the tail bound for X in a form that is easier to evaluate, and then use standard arguments (see, for example, [39, Theorem 2.3]) to derive the moment bound. We have,

$$\Pr\left(|X - \mathbb{E}[X]| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2}{2v^2}, \frac{t}{2b}\right\}\right) \tag{75}$$
$$\leq 2\exp\left(\frac{-t^2}{2(v^2 + bt)}\right), \tag{76}$$

that is,

$$\Pr\left(|X - \mathbb{E}[X]| \geq bu + \sqrt{b^2u^2 + 2v^2u}\right) \leq e^{-u}. \tag{77}$$

With this tail bound, we can now derive the stated moment bound by using

$$\mathbb{E}\left[|X - \mathbb{E}[X]|^{2q}\right] = 2q\int_0^\infty \Pr\left(|X - \mathbb{E}[X]| \geq t\right)t^{2q-1}dt. \tag{78}$$

In particular, upon substituting $t = bu + \sqrt{b^2u^2 + 2v^2u}$, we get

$$\mathbb{E}\Big[(X - \mathbb{E}[X])^{2q}\Big] \leq 2q \int_0^\infty e^{-u}(bu + \sqrt{b^2u^2 + 2v^2u})^{2q-1} \times \left(b + \frac{b^2u + v^2}{\sqrt{b^2u^2 + 2v^2u}}\right) du, \tag{79}$$

which after simplification yields

$$\mathbb{E}\Big[(X - \mathbb{E}[X])^{2q}\Big] \leq 2q \cdot (16)^q \Big(b^{2q}\Gamma(2q) + v^{2q}\Gamma(q)\Big). \tag{80}$$

□

REFERENCES

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

[2] 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.

[3] 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.

[4] L. Ramesh, C. R. Murthy, and H. Tyagi, "Sample-measurement tradeoff in support recovery under a subgaussian prior," December 2019. [Online]. Available: http://arxiv.org/abs/1912.11247

[5] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Simultaneous joint sparsity model for target detection in hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 676–680, 2011.

[6] M. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Collaborative sparse regression for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 341–354, 2014.

[7] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[8] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, "Sparse coding with anomaly detection," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[9] E. Arias-Castro, E. J. Candès, and Y. Plan, "Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism," *Ann. Statist.*, vol. 39, no. 5, pp. 2533–2556, 10 2011. [Online]. Available: https://doi.org/10.1214/11-AOS910

[10] K. Balasubramanian, K. Yu, and T. Zhang, "High-dimensional joint sparsity random effects model for multi-task learning," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'13. Arlington, Virginia, USA: AUAI Press, 2013, p. 42–51.

[11] N. Vaswani and J. Zhan, "Recursive recovery of sparse signal sequences from compressive measurements: A review," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3523–3549, 2016.

[12] J. F. C. Mota, N. Deligiannis, A. C. Sankaranarayanan, V. Cevher, and M. R. D. Rodrigues, "Adaptive-rate reconstruction of time-varying signals with application in compressive foreground extraction," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3651–3666, 2016.

[13] 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.

[14] 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.

[15] L. Ramesh, C. R. Murthy, and H. Tyagi, "Sample-measurement tradeoff in support recovery under a subgaussian prior," in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 2709–2713.

[17] Y. Wang, D. Wipf, J.-M. Yun, W. Chen, and I. Wassell, "Clustered sparse bayesian learning," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, ser. UAI'15. Arlington, Virginia, USA: AUAI Press, 2015, p. 932–941.

[16] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Ann. Statist.*, vol. 39, no. 1, pp. 1–47, 2011.

[18] Y. Qi, D. Liu, D. Dunson, and L. Carin, "Multi-task compressive sensing with dirichlet process priors," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 768–775. [Online]. Available: https://doi.org/10.1145/1390156.1390253

[19] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, p. 41–48.

[20] D. Yin, R. Pedarsani, Y. Chen, and K. Ramchandran, "Learning mixtures of sparse linear regressions using sparse graph codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1430–1451, 2019.

[21] A. Krishnamurthy, A. Mazumdar, A. McGregor, and S. Pal, "Sample complexity of learning mixture of sparse linear regressions," in *Neural Information Processing Systems*, 2019, pp. 10 531–10 540.

[22] Y. Li and Y. Liang, "Learning mixtures of linear regressions with nearly optimal complexity," in *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet, and P. Rigollet, Eds., vol. 75. PMLR, 2018, pp. 1125–1144.

[23] S. Chen, J. Li, and Z. Song, "Learning mixtures of linear regressions in subexponential time via fourier moments," *CoRR*, vol. abs/1912.07629, 2019. [Online]. Available: http://arxiv.org/abs/1912.07629

[24] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statistics and Computing*, vol. 20, no. 2, pp. 231–252, 2010.

[25] F. McSherry, "Spectral partitioning of random graphs," in *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, Oct 2001, pp. 529–537.

[26] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, p. 036104, Sep 2006.

[27] B. Hajek, Y. Wu, and J. Xu, "Semidefinite programs for exact recovery of a hidden community," *Journal of Machine Learning Research*, vol. 49, no. June, pp. 1051–1095, Jun 2016, 29th Conference on Learning Theory, COLT 2016.

[28] E. Abbe, "Community detection and stochastic block models: Recent developments," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6446–6531, 2017.

[29] L. Ramesh, C. R. Murthy, and H. Tyagi, "Multiple support recovery using very few measurements per sample: Supplementary," 2021. [Online]. Available: https://github.com/lekshmi-ramesh/SupportRecovery/blob/master/LR-CM-HT-21b-supplementary.pdf

[30] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," vol. 51, no. 2, p. 918–930, Nov. 2006. [Online]. Available: https://doi.org/10.1016/j.csda.2005.09.010

[31] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[32] T. Tao, *Topics in Random Matrix Theory*, ser. Graduate Studies in Mathematics. American Mathematical Society, 2016.

[33] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[34] M. Rudelson, "Random vectors in the isotropic position," *Journal of Functional Analysis*, vol. 164, no. 1, pp. 60 – 72, 1999.

[35] A. Chakrabarti, "Lecture notes on data stream algorithms," May 2020. [Online]. Available: https://www.cs.dartmouth.edu/~ac/Teach/CS35-Spring20/Notes/lecnotes.pdf

[36] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the Davis–Kahan theorem for statisticians," *Biometrika*, 2015.

[37] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[38] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-gaussian concentration," *Electron. Commun. Probab.*, vol. 18, p. 9 pp., 2013. [Online]. Available: https://doi.org/10.1214/ECP.v18-2865

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