Multiple Support Recovery Using Very Few Measurements Per Sample

Lekshmi Ramesh Chandra R. Murthy 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.**

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 linear measurements [18], [6], [12], [15], that has been applied to solve inverse problems in imaging, source localization, and anomaly detection [5], [7], It is also related to the study of sparse random effects in mixed linear models [3], [4]. 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. The problem of multiple support recovery is also discussed in [9], [20] under the assumption of slowly varying supports.

There are two sets of unknowns in our setting: 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 samples with the same support together, and use standard algorithms to recover the support. However, in the absence of access to the 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 [21], [1]. 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.

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 [21] 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 [18], [6], [12].

In the $m < k$ regime, on the other hand, we recently showed 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 [14], [15]. In fact, the lower bound of [15] 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 [11], the problem of recovering the union of supports from linear measurements is considered. Another

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 PhD fellowship from the Ministry of Electronics and Information Technology, Govt. of India, and by research grants from the Aerospace Network Research Consortium and the Center for Networked Intelligence (CNI) at the Indian Institute of Science.

[1]The support of $x \in \mathbb{R}^d$ is the set $\texttt{supp}(x) = \{u \in [d] : x_u \neq 0\}$.

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

line of related works is on multi-task learning/multi task sparse estimation [22], [13], [2]. 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.

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 present a sketch of the proof of our main result in Section IV. We provide the full proofs of our results, along with experiments on recovering handwritten images from random projections in the longer version of this paper.

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

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

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

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

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

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

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 $\texttt{supp}\,(X_i) \stackrel{\text{def}}{=} \{j \in [d] : X_{ij} \neq 0\} \in \{\mathcal{S}_1, \ldots, \mathcal{S}_\ell\}$, 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]

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

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 with zero mean and diagonal covariance matrix $K_{\lambda_i} = \texttt{diag}\,(\lambda_i)$, where the parameter λ_i is a d-dimensional vector for which $\texttt{supp}\,(\lambda_i) = \mathcal{S}_i$, $i \in [\ell]$. More concretely, we make the following assumption.

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

Assumption 1. *For a sample $X_j \sim \mathrm{P}^{(i)}$, $j \in [n]$, $i \in [\ell]$, and an absolute constant c, $\mathbb{E}_{\mathrm{P}^{(i)}}\left\{X_j X_j^T\right\} = \texttt{diag}\,(\lambda_i)$ with $\lambda_i \in \mathbb{R}_+^d$, $\texttt{supp}\,(\lambda_i) = \mathcal{S}_i$, 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 $\mathcal{S}_2$.

We seek to determine the value of n for which an (n, ε, δ)-estimator exists. 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 $n^*(\ell, m, k, d, \varepsilon, \delta)$.

We have the following result.

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

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

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 provide a sketch of the proof of Theorem 1 in Section IV.

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 [15] 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. Closing the gap between the lower bound and the upper bound 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. 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 projections of the data. We circumvent this difficulty by using the approach followed in [15].

Specifically, we compute $\Phi_i^\top Y_i$ and use these as a proxy for the data. We build 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 [15]. We then construct an affinity matrix using proxy samples $\Phi_i^\top Y_i$ and apply spectral clustering to estimate individual supports from the union.

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 [15] 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 $|\mathcal{S}_{\text{un}}| = k\ell$. Next, define vectors $a'_1, \ldots, a'_n$ with entries

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

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 in any support should result in a smaller value of the statistic. As shown in [15], 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 an estimate for $k\ell$, similar to the approach of using scree plots to determine the model order in principal component analysis [23].

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 $a_i \in \mathbb{R}_+^{k\ell}$. That is, $a_i = (a'_i)_{\hat{\mathcal{S}}_{\text{un}}}$. Also, without loss of generality, we set $\hat{\mathcal{S}}_{\text{un}} = [k\ell]$.[4]

[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.

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

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

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 key step here is the construction of the affinity matrix T that can provide a 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 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 the case when $\ell = 2$, 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 [16], [10], 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 algorithm in the general case is provided in Algorithm 1. Although we only have access to T, the eigenvectors of T itself suffice for clustering, provided we have sufficiently many samples. We also note that the ℓ-means step in the algorithm can be implemented using standard algorithms like Lloyd's algorithm [8].

Algorithm 1: Multiple support recovery

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

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

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

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

2 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

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

3 Restrict $a'_1, \ldots, a'_n$ to the coordinates in $\hat{\mathcal{S}}_{\text{un}}$, to get $a_1, \ldots, a_n$. 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 $\hat{\mathcal{S}}_{\text{un}}$.

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

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

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

6 *(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.

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

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

IV. Sketch of the Analysis

In this section, we provide an overview of the analysis of our estimator that leads to Theorem 1. The analysis of the first step where we recover the union of the supports follows from [15], after accounting for the fact that the samples are drawn from a mixture distribution. In particular, we have the following result as a consequence of [15, Theorem 3].

Theorem 2. *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.

For the second step, our analysis follows that of spectral clustering in the stochastic block model 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 to bound $\|T - \mathbb{E}[T]\|_{op}$ could be to first show exponential concentration around the mean for *each* entry of T. This approach however requires a more careful tail splitting method [19, Exercise 2.1.7]) since the moment generating function of each summand in (6) is unbounded. We use a result due to Rudelson [17], 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 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.

We first show that the expected value of the matrix constructed in (6) indeed has a block structure determined by the true supports. Following this, we show the next result which characterizes the number of samples required to partition the coordinates in the union into ℓ disjoint supports, such that error the criterion is met.

Theorem 3. *Let* $\nu_1 \geq \cdots \geq \nu_{k\ell}$ *denote the ordered eigenvalues of* $\mathbb{E}[T]$, *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.

In the next two lemmas, we bound the spectral quantities $\|\mathbb{E}[T]\|_{op}$ and Δ_ℓ, and $\mathbb{E}\left[\max_{i\in[n]}\|a_i\|_2^2\right]$ appearing in Theorem 3.

Lemma 1. *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 2. *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* $\mathbb{E}\left[max_{i\in[n]}\|a_i\|_2^2\right] \leq n^{\frac{2}{\log k}}\mathbb{E}\left[\|a_1\|_2^{\log k}\right]^{\frac{2}{\log k}}$.

Proof of Theorem 1. The proof of Theorem 1 now follows by combining guarantees for the union recovery step from Theorem 2 and the clustering step from Theorem 3.

We begin by applying Theorem 2 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 $\mathcal{S}_1, \ldots, \mathcal{S}_\ell$ – thus, the overall number of samples needed will be the sum of samples needed for union recovery in Theorem 2 and the sample complexity determined in our analysis below. In particular, the clustering step roughly dominates the sample complexity of our algorithm.

Next, upon substituting the bounds from Lemma 1 and Lemma 2 into Theorem 3, we see that for ε-approximate recovery of the supports it suffices to have

$$\begin{aligned} n &\geq \frac{c}{\varepsilon}\lambda_0^2\frac{k^3\ell}{m^2}\frac{\ell^2}{\lambda_0^4 k^2}\cdot n^{\frac{2}{\log k}} \\ &\quad\times\left(\lambda_0\frac{k\sqrt{k}\sqrt{\ell}}{m}(\log k)^2\right)^2\cdot\log(k\ell) \\ &= \frac{c}{\varepsilon}\frac{k^4\ell^4}{m^4}n^{\frac{2}{\log k}}(\log k)^4\log(k\ell). \end{aligned}$$

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. Discussion

In this work, we focused on the recovery of equal-sized, disjoint supports in a measurement-constrained setting. It would be interesting to extend the algorithm to handle unequal-sized, overlapping supports. While some simple heuristics work in practice, we are not aware of any theoretical results for the $m < k$ setting. Also, our work shows a sufficient condition on the number of samples required for multiple support recovery; obtaining a matching necessary condition is a challenging task in general. It requires characterizing the distance between mixture distributions, and using a component wise distance bound leads to the same lower bound as in [15] (with an additional $1/\ell$ factor).

References

[1] 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.

[2] 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.

[3] 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

[4] 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.

[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] 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.

[7] 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.

[8] S. P. Lloyd, "Least squares quantization in pcm." *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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

[14] 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.

[15] 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

[16] 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.

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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