Sample-Measurement Tradeoff in Support Recovery Under a Subgaussian Prior

Lekshmi Ramesh Chandra R. Murthy Himanshu Tyagi

***Abstract*—Data samples from $\mathbb{R}^d$ with common support of size k are accessed through m linear projections per sample. In the measurement-starved regime of $m < k$, how many samples are needed to recover the common support? We answer this question for a generative model with independent samples drawn from a subgaussian prior. We show that $n = \Theta((k^2/m^2)\log(k(d-k)))$ samples are necessary and sufficient to exactly recover the support. Our proposed sample-optimal estimator has a closed-form expression and has computational complexity of $O(dnm)$.**

I. Introduction

A set of n vectors has a common support of cardinality k that is much smaller than the dimension d of the vectors. It is easy to find this common support simply by looking at a single vector. But this will require d measurements, one for each coordinate of the vector. As is now well-known, we can make do with $m = O(k \log d/k)$ random linear measurements on a single vector by using compressive sensing based algorithms, and thereby recover its support. However, since in our setting we have multiple samples, we may use fewer measurements. Can we still recover the unknown support with k *overall* measurements using $m < k$ measurements per sample (i.e., would $nm \approx k$ suffice even when $m < k$)? We examine this question in a natural Bayesian setting and answer it in the negative: when $m < k$, we will need at least k^2/m overall measurements. Thus, in sharp contrast with the $m > k$ regime where k overall measurements suffice, a much larger number of overall measurements are necessary when $m < k$.

Specifically, consider independent samples $X_1, \ldots, X_n$ where each X_i is a zero-mean Gaussian vector of length d with covariance matrix $\texttt{diag}(\lambda)$. We assume that the entries λ_i are either 0 or 1, whereby the common support of the vectors coincides with the locations of 1s. We make linear measurements on the vectors X_i using independent random Gaussian matrices Φ_i with columns that have unit expected squared norms. The goal is to recover the common support using measurements $Y_i = \Phi_i X_i$, $1 \le i \le n$. We show that for $m < \alpha k$ with $\alpha < 1$, the minimum number of samples required to recover the support correctly with large probability is $\Theta((k^2/m^2)\log(k(d-k)))$.

The sample-optimal estimator we propose entails forming an estimate $\tilde{\lambda}$ of λ and then obtaining the support by selecting the k largest entries of $\tilde{\lambda}$. The estimate $\tilde{\lambda}_i$ has a closed-form expression: it is simply the empirical average $\frac{1}{n}\sum_{j=1}^{n}(\Phi_{ji}^{\top} Y_j)^2$ where Φ_{ji} denotes the ith column of the jth measurement matrix. This is in contrast to the standard Sparse Bayesian Learning (SBL) approach where maximum likelihood (ML) estimate is used, which can only be expressed as an (often nonconvex) optimization problem. Furthermore, the proposed estimator works under a much broader setting, one with subgaussian prior on X_1^n and subgaussian measurement matrices, and with additive subgaussian noise.

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

Information-theoretically optimal support recovery in the single sample setting is well-understood. In particular, [13] shows that for a deterministic input vector, $m = \Theta(k \log(d-k))$ measurements are necessary and sufficient to exactly recover the support using a Gaussian measurement matrix, essentially establishing that support recovery is impossible in the $m < k$ regime using a single sample. In the multiple sample setting, [10] showed a lower bound on sample complexity of support recovery of roughly (k/m), but the results are not tight for our measurement-starved regime of $m < k$. Despite several other works [12], [6], [14], the question of tradeoff between m and n when $m < k$ has not been fully addressed. Two initial works in this direction were [9] and [3], followed by [7] and [11], where it was empirically demonstrated that with multiple samples, it is indeed possible to operate in the $m < k$ regime. In another recent work closely related to ours [8], the authors demonstrate the possibility of operating in the $m < k$ regime. The error exponent, however, is

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.

expressed in terms of the eigenvalues of certain matrices and its exact dependence on the parameters k, m and d is not clear. In summary, none of these works shed light on the sample complexity of support recovery or on the tradeoff between measurements per sample and overall measurements.

Our formulation of support recovery in a Bayesian setting relates naturally to works on covariance estimation. Perhaps the closest to our setting is [2] which studies the problem of covariance estimation from low-dimensional projections of the data. However, no structural assumptions are made on the covariance matrix, and the general result in [2, Corollary 3] is loose for our setting. Also, our setting is closely related to the recently proposed inference under local information constraints setting of [1]. We impose information constraints on each sample by allowing only m linear measurements per sample.

We formulate our problem and present our main result in the next section. Our estimator and its analysis are given in Section III, and the proof of lower bound is in Section IV. We conclude with a discussion on some interesting extensions in the final section.

II. Problem formulation and main result

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

In the basic setting, the input comprises n independent samples $X_1, \ldots, X_n$ in $\mathbb{R}^d$, with each X_i having a zero-mean Gaussian distribution. We denote the covariance of X_i by $K_\lambda \stackrel{\text{def}}{=} \mathtt{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, where the d-dimensional vector λ has entries $\lambda_1, \lambda_2, \ldots, \lambda_d$ such that $\lambda \in \mathcal{S}_{k,d} \stackrel{\text{def}}{=} \left\{\lambda \in \{0,1\}^d : \|\lambda\|_0 = k\right\}$. That is, the (random) data vectors have a common support $S = \mathtt{supp}(\lambda)$ of size at most k. Each X_i is passed through a random $m \times d$ measurement matrix Φ_i, $1 \le i \le n$, with independent, zero-mean Gaussian entries, and the observations $Y_i = \Phi_i X_i \in \mathbb{R}^m$ are made available to a center. Using the measurements $Y_1, ..., Y_n$, the center seeks to determine the common support S.

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

$$\Pr\left(\hat{S}(Y^n) \neq \mathtt{supp}(\lambda)\right) \le \delta, \quad \forall\, \lambda \in \mathcal{S}_{k,d}. \qquad (1)$$

[1]Note that the value $\delta = 1/3$ is chosen here for convenience and can be replaced with any acceptable value below $1/2$.

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

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

Remark 1. Our formulation assumes that we know the support size k exactly. In fact, our proposed estimator extends easily to the setting where we only have an upper bound of k on the support size, and we seek to output a set of size k containing the support.

Our main result is the following.

Theorem 1 (Characterization of sample complexity)**.** *For $m < k/2$, the sample complexity of support recovery in the setting above is given by*

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

We provide the optimal estimator and prove the upper bound in Section III and the information-theoretic lower bound in Section IV. In fact, our proof yields a lower bound for $m < \alpha k$ for any $\alpha < 1$. Due to lack of space, we only provide proof sketches.

Our proposed estimator and its analysis applies to a much broader setting involving subgaussian priors. Recall that a random variable X is *subgaussian with variance parameter* σ^2, denoted $X \sim \text{subG}(\sigma^2)$, if $\log \mathbb{E}\left[e^{\theta(X - \mathbb{E}[X])}\right] \le \theta^2\sigma^2/2$ for all $\theta \in \mathbb{R}$. For X_1^n, we can use any prior with subgaussian distributed entries, i.e., the entries of X_i are independent and zero-mean with $\mathbb{E}\left[X_{i,j}^2\right] = \lambda_j$ for $\lambda \in \mathcal{S}_{k,d}$ and $X_{i,j} \sim \text{subG}(\lambda_j')$, where λ_j' is the variance parameter for the subgaussian random variable $X_{i,j}$. Our analysis will go through as long as $\lambda' = \Theta(\lambda)$, namely the variance and variance parameters differ only up to a constant factor.

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

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

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

III. The estimator and its analysis

We will work with the more general setting described above with subgaussian random variables, where we assume that the variance and variance parameters are of the same order. For simplicity, we assume that $X_{i,j}$ and W_i are subgaussian with variance parameter equal to their respective variances. Also, for the measurement matrix, we work with the same parameters as those for the Gaussian ensemble and set $\mathbb{E}\left[\Phi_i(u,v)^2\right] = 1/m$, $\mathbb{E}\left[\Phi_i(u,v)^4\right] = 3/m^2$ and assume that $\Phi_i(j,k)$ is subgaussian with variance parameter $1/m$. These assumptions of equality can be relaxed to order equality up to multiplicative constants.

A. The estimator

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

Formally, we consider the estimate A for the covariance matrix of X_i's given by $A = \frac{1}{n}\sum_{j=1}^n \Phi_j^\top Y_j Y_j^\top \Phi_j$. Note that A is positive semidefinite. We form an intermediate estimate $\tilde{\lambda}$ for the variance vector λ using the diagonal entries of A as follows:

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

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

$$\tilde{S} = \{(1), ..., (k)\}, \tag{2}$$

where (i) denotes the index of the ith largest entry in $\tilde{\lambda}$. Note that evaluating $\tilde{\lambda}_i$ requires $O(nm)$ steps, whereby the overall computational complexity of (naively) evaluating our proposed estimator is roughly $O(dnm)$.

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

$$\hat{\lambda}_i = \mathbb{1}_{\{\tilde{\lambda}_i \geq \tau\}}. \tag{3}$$

Unfortunately, the threshold we will choose can depend on the set S itself, making the estimator of (3) infeasible in practice. Nonetheless, we note that if $\lambda = \hat{\lambda}$, the largest k entries of $\tilde{\lambda}$ must coincide with the support of λ. Therefore,

$$\Pr\left(\tilde{S} \neq \texttt{supp}(\lambda)\right) \leq \Pr\left(\hat{S} \neq \texttt{supp}(\lambda)\right). \tag{4}$$

Using this observation, it suffices to analyze the estimator $\hat{\lambda}$ in (3), which will be our focus below.

An easy calculation shows that $\tilde{\lambda}_i$ is a biased estimate of λ_i. In particular, we have

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

where the expectation is with respect to the joint distribution of $\{X_1^n, \Phi_1^n, W_1^n\}$. We work with this biased $\tilde{\lambda}$ and account for the bias in the threshold τ.

B. And its analysis

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

We now present the performance of our estimator.

Theorem 2. *Let $\hat{\lambda}$ be the estimator described in (3), and let $\hat{S}$ be its support. Then, $\hat{S}$ equals the true support with probability at least $1-\delta$ provided*

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

for an absolute constant c.

Note that the result above applies for all k and m, and not only to our regime of interest $k < m$. Using (4), we get the same performance guarantees for our estimator (2). In particular, for $\sigma^2 = 0$, we obtain the upper bound claimed in Theorem 1.

The proof of Theorem 2 entails a careful analysis of tails of $\tilde{\lambda}_i$ and uses standard subgaussian and subexponential concentration bounds. The details are tedious and are relegated to the longer version of the paper. Below, we provide a brief outline, highlighting the main steps.

Denoting by S the support of λ and by $\hat{S}$ the support of $\hat{\lambda}$, $\Pr\left(\hat{S} \neq S\right)$ can be bounded as

$$\Pr\left(\hat{S} \neq S\right) \leq \sum_{i \in S} \Pr\left(\tilde{\lambda}_i < \tau\right) + \sum_{i' \in S^c} \Pr\left(\tilde{\lambda}_{i'} \geq \tau\right). \tag{5}$$

Our approach entails deriving tail bounds for $\tilde{\lambda}_i$, and then choosing a threshold τ to obtain the desired bound for (5); we derive lower tail bounds for $i \in S$ and upper tail bounds for $i' \in S^c$.

To that end, note

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

where we used $Y_j = \Phi_j X_j + W_j$. We proceed by observing that conditioned on Φ_1^n, $\tilde{\lambda}_i$ is a sum of independent subexponential random variables. Recall that a random variable X is *subexponential with parameters v^2 and b*, denoted $X \sim \text{subexp}(v^2, b)$, if $\mathbb{E}\left[\exp(\theta(X - \mathbb{E}X))\right] \le \exp\left(\theta^2 v^2/2\right)$, for $|\theta| < 1/b$. Using basic properties of subexponential random variables and the connection between subgaussian and subexponential random variables, we get that conditioned on measurement matrices Φ_1^n, the random variable $\tilde{\lambda}_i$ is distributed as $\text{subexp}\left(\frac{c_1}{n^2}\sum_{j=1}^n \alpha_{ji}^4, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2\right)$, where c_1 and c_2 are absolute constants and

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

The next lemma follows from standard concentration bounds for subexponential random variables.

Lemma 1. *Using our foregoing notations and denoting $\mu_i \stackrel{\text{def}}{=} \mathbb{E}\left[\tilde{\lambda}_i|\Phi_1^n\right] = \frac{1}{n}\sum_{j=1}^n \alpha_{ji}^2, i \in [d]$, for $i \in S$,*

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

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

$$\mathrm{P}(\tilde{\lambda}_{i'} \ge \tau|\Phi_1^n) \le \exp\left(-\min\left\{\frac{n^2(\tau-\mu_{i'})^2}{c_1\sum_{j=1}^n \alpha_{ji'}^4}, \frac{n(\tau-\mu_{i'})}{c_2\max_{j\in[n]}\alpha_{ji'}^2}\right\}\right).$$

Thus, using (5), we can obtain bounds for the error probability by showing that with large probability Φ_1^n takes values for which we get each term above bounded by roughly $\delta/(2\max\{(d-k), k\})$. In particular, using a manipulation of the expression for exponents, each of the conditional probability above will be less than δ if τ satisfies the following condition for any $i \in S$ and $i' \in S^c$: $\mu_{i'} + \nu_{i'} \le \tau \le \mu_i - \nu_i$, where

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

This sufficient condition can be rewritten as

$$\frac{1}{n}\sum_{j=1}^n \alpha_{ji}^2 - \frac{1}{n}\sum_{j=1}^n \alpha_{ji'}^2 \ge c'\ln\frac{1}{\delta}\left(\sqrt{\frac{1}{n^2}\sum_{j=1}^n \alpha_{ji}^4} + \sqrt{\frac{1}{n^2}\sum_{j=1}^n \alpha_{ji'}^4}\right), \quad (6)$$

where $c' = \max\{c_1, c_2\}$. Indeed, this condition holds with large probability for an appropriate choice of n.

Lemma 2. *For fixed $i \in S$ and $i' \in S^c$, (6) holds with probability at least $1-\delta$ if $n \ge c(k/m + \sigma^2)^2 \log\frac{1}{\delta}$.*

Theorem 2 follows from Lemma 2 and a union bound. Also note that the separation condition (6) fails to hold for $n = 1$, regardless of which measurement ensemble is used. This is to be expected in view of the lower bound when $m < k$, which we prove next.

IV. Sketch of proof of the lower bound

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

We consider the Bayesian hypothesis testing problem where we observe Y^n and seek to determine U. Any fixed support estimator $\hat{S}$ with probability of error less than δ will give an estimate $\hat{U}$ for U, and clearly, $\Pr\left(\hat{U} \ne U\right) = \Pr\left(\hat{S} \ne S_U\right)$, which must be less than δ. Now, by Fano's inequality and convexity of the Kullback-Leibler (KL) divergence, we get

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

where $\mathrm{P}_{Y^n|S}$ denotes the distribution of Y^n when the support of λ is S. Note that $\mathrm{P}_{Y^n|S} = \prod_{i=1}^n \mathrm{P}_{Y_i|S}$ with each $\mathrm{P}_{Y_i|S}$ having the same distribution, denoted by $\mathrm{P}_{Y|S}$. Thus, $D(\mathrm{P}_{Y^n|S_u}\|\mathrm{P}_{Y^n|S_0}) = nD(\mathrm{P}_{Y|S_u}\|\mathrm{P}_{Y|S_0})$.

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

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

$$D\left(\mathrm{P}_{Y|S_u,\Phi}\|\mathrm{P}_{Y|S_0,\Phi}\right) = \frac{1}{2}\left(\log\frac{|A_{S_0}|}{|A_{S_u}|} + \mathrm{Tr}(A_{S_0}^{-1}A_{S_u}) - m\right)$$

$$\leq \frac{1}{2}\sum_{i=1}^{m}\left(\log\frac{b_i}{a_i} - \left(1-\frac{a_i}{b_i}\right)\right)$$
$$\leq \frac{1}{2}\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_i b_i},$$

where the first inequality holds since for symmetric, positive-definite matrices A and B with eigenvalues $a_1 \geq \cdots \geq a_m$ and $b_1 \geq \cdots \geq b_m$, respectively, $Tr(AB) \leq \sum_{i=1}^{m} a_i b_i$. Using convexity of KL divergence, the Cauchy-Schwarz inequality, and the fact that a_is and b_is are identically distributed, we can get

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

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

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

By the Hoffman-Wielandt inequality [5], we have $\sum_{i=1}^{m}(a_i-b_i)^2 \leq \|A_{S_0} - A_{S_u}\|_F^2$ where the right-side coincides with $\|\Phi_1\Phi_1^\top - \Phi_{k+1}\Phi_{k+1}^\top\|_F^2$. Using the triangle inequality for Frobenius norm and noting that $\|\Phi_i\Phi_i^\top\|_F$ equals $\|\Phi_i\|_2^2$ for a vector Φ_i, we get

$$\mathbb{E}\left[\sum_{i=1}^{m}\frac{(a_i-b_i)^2}{a_m^2}\right] \leq 4\sqrt{\mathbb{E}\left[\frac{1}{a_m^4}\right]}\sqrt{\mathbb{E}\left[\|\Phi_1\|_2^8\right]}$$
$$\leq c'\sqrt{\mathbb{E}\left[\frac{1}{a_m^4}\right]},$$

where the last inequality uses the expression for the fourth moment of a chi-square random variable.

It only remains to bound $\mathbb{E}\left[1/a_m^4\right]$, where a_m is the smallest eigenvalue of the $(m \times m)$ Wishart matrix A_{S_u}. By using a tail-bound for the minimum eigenvalue of a Wishart matrix [4, Lemma 4.1], we get

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

By combining all the steps above, we get

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

for a constant c, which yields the desired bound.

V. Discussion

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

First, one can consider using the same measurement matrix for all samples. In this case, we observe empirically that our estimate does not perform well, but we do not have a theoretical understanding of this phenomenon (see [2, Proposition 2] for a related discussion). Next, our current results are tight only for the high SNR case of $\sigma^2 < k/m$; it will be of interest to get optimal bounds for all σ^2. Also, it would be interesting to extend our results to nonbinary λ, where only a lower bound is assumed for nonzero entries of λ. Finally, one can study the problem of support recovery when k is unknown.

Acknowledgment

We thank Manjunath Krishnapur for educating discussions on moments of eigenvalues of a Wishart matrix.

References

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

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

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

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

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

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

[7] S. Khanna and C. R. Murthy, "Rényi divergence based covariance matching pursuit of joint sparse support," in *SPAWC*, 2017, pp. 1–5.

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

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

[10] 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.

[11] L. Ramesh and C. R. Murthy, "Sparse support recovery via covariance estimation," in *ICASSP*, 2018, pp. 6633–6637.

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

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