Support Recovery from Linear Measurements: Tradeoffs in the Measurement-Constrained Regime

A Thesis

Submitted for the Degree of

Doctor of Philosophy

in the **Faculty of Engineering**

by

Lekshmi Ramesh

under the Guidance of

Prof. Chandra R. Murthy
and
Prof. Himanshu Tyagi

Electrical Communication Engineering
Indian Institute of Science
Bangalore – 560 012, INDIA

March 2021

TO

My family.

Acknowledgments

I have been fortunate to have *Prof. Chandra R. Murthy* and *Prof. Himanshu Tyagi* as my advisors. It has been an inspiring experience for me to have had the opportunity to learn from both of them. I thank them for being extremely kind and patient especially in the initial years, for always asking insightful questions, and for being available for discussions despite their busy schedules. The guidance and encouragement I have received from both of them during my PhD has defined my research, and it was a pleasure to have been their student.

My years at IISc would certainly have been dull if it was not for the wonderful bunch of friends I had there. Thanks to *Purvi, Shubham, Zaki*, and *Pushtivardhan*, whom I have known from the time of my Masters, for all the good memories, random discussions during food, birthday celebrations, and most of the time, for just being there. I would like to thank the members of the *SPC Lab*: *Saurabh, Ribhu, Mohit, Geethu, Pradip, Sai, Arun, Shilpa, Sireesha, Chethan, Monica, Akshay, Chirag, Anubhab, Soumendu* and everyone else that I have had the chance to interact with during my time there. Thanks especially to *Saurabh, Sai* and *Pradip* for the many useful discussions in and outside the lab, and for being great labmates. Thanks also to all my other friends from the department: *Vinay, Garima, Sahasranand, Raghava, Prathamesh*, and *Karthik*. Finally, much thanks to *Avni* for hosting me in her PG room on multiple occasions, and to my amazing group of school friends for bringing in some much needed joy and cheer every once in a while.

Last year was an anomaly no one could have predicted, with nearly everything coming to a standstill due to the pandemic. Like everyone else, I spent more than a year working from home. I can't thank my parents and my sister enough for always supporting me,

and especially for being understanding and patient during this time. None of this would have been possible without their love and sacrifice.

Publications and Preprints

1. Lekshmi Ramesh, Chandra R. Murthy, and Himanshu Tyagi, "Phase Transition for Support Recovery from Gaussian Linear Measurements", `https://arxiv.org/abs/2102.00235` *[cs.IT], February, 2021; Under review at ISIT 2021.*

2. Lekshmi Ramesh, Chandra R. Murthy, and Himanshu Tyagi, "Multiple Support Recovery Using Very Few Measurements Per Sample", *February, 2021; Under review at ISIT 2021.*

3. Lekshmi Ramesh, Chandra R. Murthy, and Himanshu Tyagi, "Multiple Support Recovery Using Very Few Measurements Per Sample", *February, 2021; Longer version, under review at IEEE Transactions on Signal Processing.*

4. Lekshmi Ramesh, Chandra R. Murthy, and Himanshu Tyagi, "Sample-Measurement Tradeoff in Support Recovery under a Subgaussian Prior", `https://arxiv.org/abs/1912.11247` *[cs.IT], December, 2019; Longer version, under review at IEEE Transactions on Information Theory.*

5. Lekshmi Ramesh, Chandra R. Murthy, and Himanshu Tyagi, "Sample-Measurement Tradeoff in Support Recovery under a Subgaussian Prior", *IEEE International Symposium on Information Theory (ISIT), Paris, Jul. 2019, pp. 2709–2713.*

6. Lekshmi Ramesh and Chandra R. Murthy, "Sparse Support Recovery via Covariance Estimation", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, Canada, Apr. 2018, pp. 6633–6637.*

Abstract

In this thesis, we study problems under the theme of discovering joint structure in a set of high-dimensional data samples from linear measurements. Our primary focus is on the regime where the number of samples available is large, but we are constrained to access very few measurements per sample. This setting can be used model high dimensional estimation tasks in a distributed setting, where storing or communicating more measurements per sample can be expensive. We study a basic problem in this setting – that of support recovery from linear measurements. In this problem, a set of n samples in $\mathbb{R}^d$, each having a support of size k, is accessed through m linear measurements per sample. The goal is to recover the unknown support, given knowledge of the measurements and the measurement matrices. This problem, also sometimes referred to as variable selection or model selection, has been extensively studied in the signal processing and statistics literature, and finds applications in source localization, hyperspectral imaging, heavy hitters detection in networks, and feature selection in regression. It is known that if we have $m = \Omega(k \log d)$ measurements per sample, then a single *sample is sufficient for support recovery. As such, when we have access to multiple samples, an interesting question to consider is whether we can perform recovery with $m < k$ measurements per sample. This measurement-constrained setting is relatively less explored in the literature, and the optimal sample-measurement tradeoff was unknown prior to our work.*

We provide a tight characterization of the sample complexity of this problem, which together with previous results in the literature gives a full understanding of this problem for all values of k/m. We propose two algorithms that can perform recovery in the

measurement-constrained regime, where standard algorithms fail to work. Our first algorithm is a simple, closed-form variance estimation-based procedure, while our second algorithm is based on an approximate maximum likelihood procedure. We show that when $m < k$, the minimum number of samples required for exact support recovery with high probability scales as $\Theta((k^2/m^2)\log d)$, and the closed-form estimator achieves this scaling.

To obtain the upper bound on sample complexity, we analyze the closed-form estimator for both random inputs drawn from a subgaussian prior, and for deterministic, worst-case inputs. We show that in either case, the upper bound has the same scaling with respect to the problem dimensions. In our analysis for the worst-case input setting, we provide some useful results in the form of concentration bounds for heavy-tailed random variables, which may find use in other problems as well.

Our lower bound construction uses Gaussian samples and Gaussian measurement matrices, and is based on characterizing the distance between pairs of competing output distributions resulting from linear measurements from samples whose supports are close. The lower bound that we obtain with Gaussian inputs implies a lower bound for the deterministic inputs case as well. In fact, it matches the upper bound that we obtain for the deterministic input case, in turn showing that the case of Gaussian inputs is the hardest for the common support recovery problem. In summary, our results settle the question of tradeoff between m and n in the $m < k$ regime, and show that there exists a phase transition for the sample complexity of this problem at $k/m = 1$. Roughly, around this point, the sample complexity for support recovery undergoes a change from being linear in the ratio k/m to being quadratic in k/m (up to a factor of $\log d$).

We then consider an extension of the common support recovery problem to the case of multiple disjoint supports, where the support of each sample is assumed to be one out of a small set of ℓ allowed supports, each of size k. We propose a two-step algorithm for this setting, that first estimates the union of the underlying supports, and then estimates the individual supports using a spectral algorithm. In effect, the first step utilizes second order statistics of the data to recover the union, while the second step uses fourth order

statistics to cluster coordinates in the union into ℓ different supports. We analyze this algorithm for the class of subgaussian inputs and measurement matrices, and show an upper bound of $\tilde{O}(k^4\ell^4/m^4)$ on the sample complexity of this problem when $m < k$.

Contents

List of Figures

Notation

Sets

$\mathbb{R}$	The set of real numbers
$\mathbb{N}$	The set of natural numbers
$[n]$	The set $\{1, \ldots, n\}$
$\binom{[d]}{k}$	The set of subsets of $[d]$ of cardinality k
$\mathcal{S}_1 \times \mathcal{S}_2$	Cartesian product of the sets $\mathcal{S}_1$ and $\mathcal{S}_2$
$\mathcal{S}\Delta\mathcal{S}'$	Symmetric difference between sets $\mathcal{S}$ and $\mathcal{S}'$, i.e., $(\mathcal{S}\backslash\mathcal{S}')\cup(\mathcal{S}'\backslash\mathcal{S})$

Vectors and matrices

I_d	Identity matrix of size $d \times d$
A^T	Transpose of matrix a A
$\lvert A\rvert$	Determinant of matrix a A
$\texttt{rank}(A)$	Rank of a matrix A
$A_{\mathcal{S}}$	Submatrix of A obtained by restricting to columns indexed by the set $\mathcal{S}$
$A \succcurlyeq B$	$A - B$ is positive semi-definite, where A and B are symmetric matrices
$\lVert A\rVert_{op}$	Operator norm of matrix A, i.e. $\sup_{\lVert x\rVert_2=1} \lVert Ax\rVert_2$
$\lVert A\rVert_F$	Frobenius norm of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, i.e., $\sum_{i=1}^{d_1}\sum_{j=1}^{d_2} A_{ij}^2$
$\texttt{Tr}(A)$	Trace of a matrix A
$\texttt{vec}(A)$	Vectorized version of a matrix A

Vectors and matrices

$A \odot B$	Khatri-Rao product of matrices A and B
$A \otimes B$	Kronecker product of matrices A and B
$\texttt{diag}\,(a)$	A diagonal matrix with the entries of vector a on the diagonal
$\|a\|_p$	ℓ_p norm of a vector $a \in \mathbb{R}^d$, $p \in [1, \infty]$, i.e., $\sqrt{\sum_{i=1}^d a_i^2}$
$\texttt{supp}\,(a)$	Support of the vector $a \in \mathbb{R}^d$, i.e., the set $\{i \in [d] : a_i \neq 0\}$

Random Variables and Events

iid	Independent and identically distributed
$\Pr(\mathcal{E})$	Probability of an event $\mathcal{E}$
$\mathbb{E}[X]$	Expectation of a random variable X
$\texttt{var}(X)$	Variance of a random variable X
$\|X\|_{\mathcal{L}_p}$	$\mathcal{L}_p$ norm of a random variable X, $p \in [1, \infty)$, i.e., $(\mathbb{E}[X^p])^{\frac{1}{p}}$
$I(X;Y)$	Mutual information between random variables X and Y
$D(\mathrm{P}\|\mathrm{Q})$	Kullback-Leibler divergence between distributions P and Q
$X \sim \mathcal{N}(\mu, \sigma^2)$	X is a Gaussian random variable with mean μ and variance σ^2
$X \sim \mathrm{subG}(\sigma^2)$	X is a subgaussian random variable with parameter σ^2, i.e., $\log \mathbb{E}\left[e^{\theta(X-\mathbb{E}[X])}\right] \leq \theta^2\sigma^2/2$, for all $\theta \in \mathbb{R}$
$X \sim \mathrm{subexp}(v^2, b)$	X is a subexponential random variable with parameters v^2 and $b > 0$, i.e., $\log \mathbb{E}\left[e^{\theta(X-\mathbb{E}[X])}\right] \leq \theta^2 v^2/2$, for all $\|\theta\| < 1/b$

Order notation

$f(n) = O(g(n))$	$\exists c > 0$ such that $f(n) < cg(n)$ for all sufficiently large n
$f(n) = \Omega(g(n))$	$\exists c > 0$ such that $f(n) > cg(n)$ for all sufficiently large n
$f(n) = \Theta(g(n))$	$\exists c > 0$ such that $f(n) = cg(n)$ for all sufficiently large n

Acronyms

ML	Maximum Likelihood
MSE	Mean Squared Error
PCA	Principal Components Analysis
SBM	Stochastic Block Model
SBL	Sparse Bayesian Learning
MGF	Moment Generating Function

Miscellaneous

$\Gamma(u)$	Gamma function at $u > 0$, i.e. $\int_0^\infty x^{u-1}e^{-x}dx$

Chapter 1

Introduction

Modern applications involving high-dimensional data have led to a resurgence of interest in several classical problems in the theory of estimation and detection. Unlike the classical setting, however, the focus has shifted to the regime where the number of observations available is much smaller than the dimension of the data. This shift can be attributed to the limitations that are associated with collecting and processing high-dimensional data. On the other hand, making meaningful inference is still possible in such cases because the data typically has additional structure, which can be exploited by the inference procedure.

As a concrete example, consider the task of linear regression and the performance that the least squares estimator achieves. In particular, assume we are given observations $Y = \Phi x + W \in \mathbb{R}^m$, where $x \in \mathbb{R}^d$ are the regression coefficients, $\Phi \in \mathbb{R}^{m \times d}$ is the matrix of covariates, and $W \sim \mathcal{N}(0, \sigma^2 I)$ is additive noise. When $m \geq d$, the regression coefficients can be estimated using the least squares estimator $\hat{x}_{\mathrm{LS}} \stackrel{\mathrm{def}}{=} (\Phi^\top \Phi)^{-1} \Phi^\top Y$, and the resulting mean squared error (MSE) can be explicitly characterized. Specifically, a simple calculation shows that the MSE as a function of Φ is given by $\mathbb{E}\left[\|\hat{x}_{\mathrm{LS}} - x\|_2^2 | \Phi\right] = \sigma^2 \mathtt{Tr}((\Phi^\top \Phi)^{-1})$. When Φ has independent, standard normal entries, the MSE averaged over Φ simplifies to $O(d\sigma^2/m)$. This in turn means that as long as the number of observations scales at least linearly with the dimension d, the MSE remains bounded above by a constant.

On the other hand, when $m < d$, it is impossible to estimate x without further assumptions. If, for instance, most of the coefficients in x are expected to be zero (implying

that only a few columns of the covariates matrix are active), then a constrained version of the least squares estimator can instead be used. In particular, if at most $k < d$ coefficients are known to be nonzero, then one can perform least squares restricted to the set $\{z \in \mathbb{R}^d : |\texttt{supp}\,(z)\,| \leq k\}$, provided $k \leq m$. Unfortunately, unlike the previous case, the solution is not available in closed form, but it can be shown that for Gaussian design matrices as before, the MSE scales as $O(\sigma^2 k \log(d/k)/m)$, using a more involved analysis compared to the unconstrained version [63]. Note that in this case fewer observations are sufficient to guarantee that the MSE is small.

Although the discussion above was limited to a specific estimator and its variants, the performance bounds that we saw can be shown to be optimal [63]. Nonetheless, at a high level, the example illustrates two key points. The first is that fewer observations suffice when the data has additional structure, as might be expected. The second is that new estimators and techniques are required that exploit this underlying structure. These observations have been made in the context of several other problems as well, including support recovery, covariance estimation, matrix completion, and principal components analysis (PCA), in similar high-dimensional settings. In each of these problems, additional structural assumptions such as sparsity and low-rankness lead to error rates that scale only sublinearly with the dimension d, see [81], [14], [17], [93], [24], [10] for more details and precise statements.

In spite of a large body of work on these problems, we now point to a canonical setting that is relatively less understood, which we will refer to as the multiple sample measurement-constrained setting (or simply, the measurement-constrained setting). In this setting, there are multiple data samples with some joint structure that we wish to infer using observations made per data sample (for example, in the form of linear measurements, quadratic measurements, or random subsamples). The question that we ask is whether such inference can be done with *fewer* observations per sample than what is dictated by the single sample setting. In other words, what is the tradeoff between the number of samples and the number of measurements per sample? Understanding this tradeoff can be useful for settings where obtaining more measurements per sample

is expensive, and we would like to reduce the number of measurements collected per sample as much as possible. Our goal in this work will be to study this question in the context of the problem of support recovery from linear measurements. We will do so by characterizing the sample complexity of this problem in the measurement-constrained regime. Our results will show that the sample complexity undergoes a change as we move from the measurement-rich regime to the measurement-constrained regime, with more samples being required in the latter case. We describe the problem setting in more detail in the next section.

1.1 Support recovery from multiple samples

In the problem of support recovery from linear measurements, a set of d-dimensional data samples $x_1, \ldots, x_n$ are observed through linear measurements of the form $Y_i = \Phi_i x_i + W_i$, where $\Phi_i \in \mathbb{R}^{m \times d}$ with $m < d$ is the ith measurement matrix and W_i is additive noise, independent across samples. Each sample has a support of size $k < d$, and we will be interested in two kinds of settings. The first one is where the support is common across samples, i.e., $\texttt{supp}\,(x_i) = \mathcal{S}$ for all $i \in [n]$ for some unknown $\mathcal{S} \subset [d]$. The second case is where the support can vary, but it is drawn from a small unknown set of ℓ allowed supports. That is, for each $i \in [n]$, $\texttt{supp}\,(x_i) \in \{\mathcal{S}_1, \ldots, \mathcal{S}_\ell\}$ with $\ell \ll n$, implying that there will be multiple samples with the same support. Note, however, that the label associating samples to their respective supports is unknown. Our goal in either case is to recover the underlying support(s) with high probability. This setting has been used to model problems in hyperspectral imaging, source localization, anomaly detection, and mixed linear models [20], [35], [44], [3], [6], [8].

We will first focus on the case of a common support to describe what was known prior to our work and to highlight our results. As we will describe in detail in subsequent chapters, the $n = 1$ case, where there is a single unknown sample, is fairly well-understood. It is known that $m = \Theta(k \log(d-k))$ measurements are necessary and sufficient to recover $\mathcal{S}$ exactly with high probability using a Gaussian measurement matrix [81](see the next chapter for a more precise statement involving conditions on the SNR). Following this, it

was shown that when multiple samples are available, i.e., when $n > 1$, $mn = \Theta(k \log(d/k))$ *overall measurements*[1] are necessary and sufficient provided $m > k$ [55]. The upper bound from [55] uses an exhaustive search decoder, but a similar condition of $m > k$ is required for other estimators such as $\ell_{1,2}$ minimization as well [51], and we call this the *measurement-rich* regime. This suggests that regardless of how large n is, we still require at least k measurements per sample (which is roughly what is required in the $n = 1$ case). Furthermore, in this measurement-rich regime, the two resources m and n have a similar effect on the recovery performance, and only the overall number of measurements mn matters.

However, when we have multiple samples with joint structure (e.g., common or repeating supports), it is natural to expect that we should be able to perform support recovery with fewer than k measurements per sample.[2] We will call this the *measurement-constrained* regime. For both the problems described in the beginning of this section, we will design estimators that can reliably recover the support(s) in the measurement-constrained regime. For the problem of a common support, we will also show the optimality of our proposed estimator through a lower bound on the sample complexity. Our results show that compared to the $m > k$ regime, more samples are required in the $m < k$ regime, and the effect of the parameters m and n on the recovery performance is different. In summary, our results, together with previous results in the literature for the measurement-rich regime provide a full understanding of the sample complexity of this problem for all regimes of m and k. For the setting with multiple supports, we will derive an upper bound on the sample complexity by analyzing a spectral clustering-based algorithm. We will also demonstrate the empirical performance of our algorithm on synthetic and real datasets, and show that it can perform support recovery with very few measurements per sample.

[1] We refer to the quantity mn as the *overall* number of measurements in our model.

[2] Clearly, support recovery with $m < k$ measurements is impossible using a single sample as dictated by the lower bound for $n = 1$.

1.2 Summary of contributions and techniques

In this section, we briefly describe the contributions of this thesis.

1. We characterize the sample complexity $n^*(m, k, d)$ of support recovery when there is a single unknown support, and show that in the measurement-constrained regime of $m < k$, $n^*(m, k, d) = \Theta((k^2/m^2) \cdot \log d)$. Our results thus demonstrate a change in the behavior of the sample complexity from linear to quadratic in the ratio k/m (up to logarithmic factors) as we move from the $m > k$ to the $m < k$ regime. We also show that a closed-form estimator based on estimating coordinate-wise variances achieves the optimal scaling. See Theorem 3.2.1for a statement of the result.

2. Our sample complexity bound holds for both random and deterministic inputs. In particular, we analyze the performance of the closed-form estimator for inputs drawn from certain subgaussian priors and for deterministic, worst-case inputs. We show that in either case, the sample complexity upper bound has the same scaling with respect to the problem dimensions. The proofs of both of our upper bounds rely on deriving concentration inequalities for heavy-tailed random variables, which in our case are functions of the measurement matrices. Standard approaches based on controlling the moment generating function (MGF) cannot be used, since the MGF is unbounded in this case. In proving the upper bounds, we will derive exponential tail bounds for quadratic forms of random vectors with heavy-tailed entries using a moment based method. See Theorems 2.3.6 and 2.4.1 for a statement of our results.

3. Our lower bound is derived using Gaussian inputs, stated in Theorem 3.1.1, and relies on characterizing the distance between pairs of output distributions resulting from linear measurements made on inputs with supports that are close to each other. The lower bound that we obtain for the Gaussian case implies a lower bound for the deterministic input setting as well. In fact, it matches the upper bound that we obtain for the deterministic input case, in turn showing that the case of Gaussian inputs is the hardest for the common support recovery problem. Our lower bound proof is based on controlling the distance between the covariance matrices of

competing output distributions measured in terms of their spectrum, and involves characterizing the expected value of quantities that are a function of the spectrum of the measurement matrices. In this respect, our lower bound proof differs from those in previous works, which proceed by controlling the difference $\|\Phi_{\mathcal{S}} x_{\mathcal{S}} - \Phi_{\mathcal{S}'} x_{\mathcal{S}'}\|_2$ between the means of the output distributions for nearby supports $\mathcal{S}$ and $\mathcal{S}'$.

4. For the case of multiple unknown supports, we propose a spectral algorithm and analyze it to obtain an upper bound (stated in Theorem 4.2.1) on the sample complexity of the multiple support recovery problem, focusing again on the measurement-constrained regime. The algorithm first computes the union of the underlying supports by using the closed-form estimator, and then obtains individual support estimates from this union estimate by performing spectral clustering on a certain matrix that depends on fourth order statistics of the inputs. Our analysis is based on characterizing the distance between the eigenvectors of this random clustering matrix from those of the expected clustering matrix.

5. Our estimators for both the single and multiple support recovery problems are based on the idea that higher order statistics of the data can reveal finer structure. We use second order statistics of the data for detecting coordinates with nonzero entries, and fourth-order statistics to further cluster the coordinates (which leads to recovery of multiple supports). In particular, the construction of the estimator for both problems is based on the idea of first forming proxy samples using the linear measurements, and then using second and fourth order statistics of these proxy samples. Alternatively, one can also view this as forming the sample mean and covariance after an initial preprocessing step which involves squaring the entries of the proxy samples. This squaring step makes the estimator robust to "cancellations" that can occur when averaging across samples that can lead to missed detection errors.

1.3 Organization of the thesis

We describe the problem of support recovery with a single unknown support in Chapter 2. Our main result is a tight characterization of the sample complexity of this problem. We describe two estimators for support recovery and prove an upper bound on the sample complexity by analyzing one of the estimators. Chapter 3 is devoted to the proof of the sample complexity lower bound. We also discuss some consequences of our results including the change in the sample complexity of support recovery as we move from the measurement-rich to the measurement-constrained regime. In Chapter 4, we consider the setting with multiple unknown supports. We describe an algorithm for this setting and derive an upper bound on the sample complexity of the multiple support recovery problem. We conclude with a discussion of possible extensions of our work in Chapter 5.

Chapter 2

Recovering a Single Support: Estimators

In this chapter, we study the problem of recovering the common k-sized support of a set of n samples of dimension d, using m noisy linear measurements per sample. Most prior work has focused on the case when m exceeds k, in which case n of the order $(k/m)\log(d/k)$ is both necessary and sufficient. Thus, in this regime, only the total number of measurements mn across the samples matter, and there is not much benefit in getting more than k measurements per sample. In the measurement-constrained regime where we have access to *fewer* than k measurements per sample, we show an upper bound of $O((k^2/m^2)\log d)$ on the sample complexity for successful support recovery when $m \geq 2\log d$. We will see two estimators that can perform recovery in the $m < k$ regime: the first is an approximate maximum likelihood (ML) estimator, and the second is a closed-form estimator. The first estimator uses the specific form of the covariance matrix resulting from linear measurements, and a Gaussian approximation step to then find the ML estimate using a nonnegative quadratic program. We empirically evaluate the performance of this estimator and show that it successfully recovers the support in the $m < k$ regime. Analyzing this estimator turns out to be difficult in general due to the fact that the ML cost is a complicated function of higher order products of the measurement

The work in this chapter is based on [58], [59], [62].

matrix. We then consider the closed-form estimator which is more amenable to analysis, and is in fact sample optimal. Our sample complexity upper bound will thus be obtained using the closed-form estimator, which we will analyze for both random and deterministic inputs.

2.1 Introduction

The problem of support recovery in the single sample setting considers the following question: given noisy linear measurements $Y = \Phi x + W \in \mathbb{R}^m$ of a k-sparse vector $x \in \mathbb{R}^d$, can we recover the locations of its nonzero entries when $m < d$? The set of indices corresponding to the nonzero entries of x is called the support of x, and is denoted by $\texttt{supp}(x)$. The measurement matrix $\Phi \in \mathbb{R}^{m \times d}$ is a design parameter that is chosen to enable exact or approximate recovery of $\texttt{supp}(x)$, and $W \sim \mathcal{N}(0, \sigma^2 I)$ is noise. This problem (also sometimes referred to as model selection or variable selection) has received a lot of attention in the past decade [81], [27], [4], [64], [47], with a focus on designing recovery algorithms and on determining the number of measurements m required for successful recovery. In particular, it is known that $m = \Theta(k \log(d-k))$ measurements are necessary and sufficient for support recovery with high probability using a Gaussian measurement matrix [81]. It is important to note that this tight scaling holds in the low signal to noise ratio (SNR) regime of $x_{\min}/\sigma^2 = \Theta(1/k)$, where $x_{\min} \stackrel{\text{def}}{=} \min_{i \in [d]} x_i$. In other regimes of SNR, either the log dependence changes or the upper and lower bounds are known to differ by a factor of $(\log(1 + k x_{\min}^2/\sigma^2))^{-1}$; see [84], [91], [47] for more discussion.

Parallel to the results in the single sample setting, there has been work on the natural extension of this problem to the multiple sample setting, which is also the focus of our work. In this setting, there are multiple samples $x_1, \ldots, x_n$, all sharing a common unknown support $\mathcal{S}$ of cardinality k. For each sample x_i, we observe measurements $Y_i = \Phi_i x_i + W_i$, and the goal is to recover $\mathcal{S}$. We can ask the question of how the number of measurements per sample m and the number of samples n can be traded off for each other, and whether it is useful to take more samples or more measurements per sample.

While there have been several works in the multiple sample setting [88], [72], [28], [36],

[69], [55], they focus on the regime where one has access to roughly $m \geq k$ measurements per sample. In particular, omitting the dependence on SNR, [55] shows that $mn = \Theta(k \log(d/k))$ is necessary and sufficient assuming $m = \Omega(k)$ and $k = o(d)$. While the sufficient condition in [55] is obtained via the analysis of an exhaustive search decoder, algorithms such as the group LASSO also show a similar scaling of $mn = \Theta(k \log(d-k))$ provided $m > k$ [51].

From the discussion in the previous paragraph, it is clear that if we have $m = \Omega(k \log(d-k))$, then a single sample is sufficient for support recovery. Therefore, given that we have access to multiple samples now, a more interesting question to consider is whether we can perform recovery with $m < k$ measurements per sample. This measurement-constrained regime has received some attention in the past [9], [54], but a characterization of the sample complexity was not known prior to our work.

In the first part of this chapter, we will show that for the case of *random* inputs drawn from a subgaussian distribution, the sample complexity upper bound (ignoring noise variance and parameters dependent on the generative model for the samples) is $n = O((k^2/m^2) \log d)$ for $(\log k)^2 \leq m < k$. In the next part, we will focus on the case of *deterministic* inputs and show that the tradeoff identified for subgaussian inputs holds for the worst-case setting as well. This result, together with our lower bound result from Chapter 3 will provide a tight characterization of the sample complexity of support recovery.

2.2 Prior work

Information-theoretically optimal support recovery in the single sample setting is well-understood and [81], [27], [4] were some of the first works to look at this problem. In particular, [81] shows that for a deterministic input vector, $m = \Theta(k \log(d-k))$ measurements are necessary and sufficient to exactly recover the support using a Gaussian measurement matrix, establishing that support recovery is impossible in the $m < k$ regime using a single sample. Following these works, several papers proposed algorithms for the multiple sample setting, that include convex programming methods [44], [75], [26], thresholding-based

methods [28], [31], Bayesian methods [88] and greedy methods [77], [25] [79]. Recovery of the support is important in several practical applications including spectrum sensing [71] and group testing [83]. Moreover, in settings where there are *multiple* unknown supports, the common support recovery algorithm can be used to estimate the union of the supports. This can be useful in problems of linear regression where there are multiple unknown subsets of correlated variables [51]. When $m \geq k$, support recovery implies recovery of the data vectors also. Indeed, given the support, one can estimate the data vectors by solving a least squares problem restricted to the support. In this work, we show that when $m < k$, support recovery is still possible. Clearly, recovery of the data vectors is no longer possible, since there are infinitely many solutions even after restricting to the support.

A setup similar to ours was studied in [55], but the results are not tight in the $m < k$ regime. In particular, [55] showed a lower bound on sample complexity of support recovery of roughly (k/m), much weaker than our $(k/m)^2$ lower bound. Another related line of works [72], [36] studies this problem considering the *same* measurement matrix for all samples, under the assumption that the data vectors are deterministic. However, none of these works characterize the tradeoff between m and n when $m < k$.

Initial works considering the $m < k$ regime were [54] and [9], followed by [37] and [59], where it was empirically demonstrated that when multiple samples are available, it is possible to operate in the $m < k$ regime. We note that the estimator we consider is similar to the one in [40]. However, the analysis in [40] is conditioned on the measurement matrix ensemble, and the error probability is expressed in terms of quantities dependent on the measurement matrix. As such, the final dependence of n on m, k and d cannot be inferred from this result. In this work, we overcome these shortcomings by showing that the estimator successfully recovers the support for a large class of subgaussian measurement matrix ensembles, and we explicitly characterize the dependence of n on m, k and d. We also provide matching lower bounds, which shows the optimality of the estimator. Two other related works that consider the $m < k$ setting are [54] and [39]. However, the precise characterization of sample complexity is not addressed in any of these works.

We will consider the case of both random and deterministic inputs, and our formulation in the random setting naturally relates to some of the works on covariance estimation. A recent work which looks at the problem of covariance estimation from low-dimensional projections of the data is [7]. As we will see in the next section, support recovery in a Bayesian setting amounts to estimating a diagonal and sparse covariance matrix, and the general result in [7, Corollary 3] for this specific case is loose and does not give the correct scaling for the sample complexity. Two other works that study covariance estimation from projected samples in the $m = 1$ case are [15] and [21]. However, these results also do not give the correct scaling on the number of samples, when applied to the diagonal sparse case. Further, since m is set to one, the tradeoff between m and n is not clear.

Our setting is also related to the recently considered inference under local information constraints setting of [2]. We impose information constraints on each sample by allowing only m linear measurements per sample. Roughly, our results say that when local information constraints are placed (namely, $m < k$), support recovery requires much more than k overall measurements.

Before moving further we make an additional remark about notation.

Remark 2.2.1 (Elaboration on notation.)**.** *We use upper case letters to denote random variables (scalars, vectors or matrices) and deterministic matrices, and lowercase letters to denote deterministic scalars or vectors. For a matrix A_j, A_{ji} denotes its ith column, $A_j(u, v)$ denotes its (u, v)th entry and $(A_j)_S$ denotes the submatrix formed by columns indexed by S. Also, for a vector X_j, X_{ji} denotes the ith component of X_j.*

2.3 The case of random inputs

We start by considering a Bayesian formulation for support recovery, where the input comprises n independent samples $X_1, \ldots, X_n$ in $\mathbb{R}^d$, with each X_i having a zero-mean Gaussian distribution. We denote the covariance of X_i by $K_\lambda \stackrel{\text{def}}{=} \texttt{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, where the d-dimensional vector λ has entries $\lambda_1, \lambda_2, \ldots, \lambda_d$, such that $\lambda \in \mathcal{S}_{k,d} \stackrel{\text{def}}{=}$

$\left\{u \in \{0,1\}^d : |\mathtt{supp}(u)| = k\right\}$. That is, the (random) data vectors have a common support $\mathcal{S} = \mathtt{supp}(\lambda)$ of size k, almost surely[1].

Each X_i is passed through a random $m \times d$ measurement matrix Φ_i, $1 \leq i \leq n$, with independent, zero-mean Gaussian entries with variance $1/m$, and the observations $Y_i = \Phi_i X_i + W_i \in \mathbb{R}^m$ are made available to a center, where the noise W_i has independent, zero-mean Gaussian entries with variance σ^2, independent of X_i and Φ_i. Using the measurements $Y_1, \ldots, Y_n$, the center seeks to determine the common support $\mathcal{S}$ of $X_1, \ldots, X_n$. To that end, the center uses an estimate $\hat{\mathcal{S}} : \mathbb{R}^{m \times n} \to \binom{[d]}{k}$, 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 δ, namely [2]

$$\Pr\left(\hat{\mathcal{S}}(Y^n) \neq \mathtt{supp}(\lambda)\right) \leq \delta, \quad \forall\, \lambda \in \mathcal{S}_{k,d}. \tag{2.1}$$

This is similar to the nonuniform recovery guarantee in compressed sensing [30, Section 9.2].

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

Definition 2.3.1 (Sample complexity of common support recovery)**.** *For $m, k, d \in \mathbb{N}$, the sample complexity of common support recovery $n^*_{\mathtt{C,avg}}(m,k,d)$ is defined as the minimum number of samples n for which we can find an estimator $\hat{\mathcal{S}}$ satisfying* (2.1) *for some $\delta \in (0,1)$, i.e.,*

$$n^*_{\mathtt{C,avg}}(m,k,d) \stackrel{\text{def}}{=} \min\left\{n : \exists \hat{\mathcal{S}} \text{ s.t. } \Pr\left(\hat{\mathcal{S}}(Y^n) \neq \mathtt{supp}(\lambda)\right) \leq \delta,\ \forall\, \lambda \in \mathcal{S}_{k,d}\right\}.$$

We will show the following upper bound on the sample complexity.

[1] Throughout, we will be considering inputs with support size *exactly* k, also sometimes referred to as exact k-sparsity.

[2] We will usually choose $\delta = 1/3$ in our upper bound results for convenience and it can be replaced with any acceptable value below $1/2$. We also note that although our upper bound results capture the dependence on δ, the lower bound which we derive later is stated for a fixed δ.

Theorem 2.3.1. *For* $(\log k)^2 \leq m < k$ *and* $1 \leq k \leq d-1$*, the sample complexity of common support recovery in the noiseless case satisfies*

$$n^*_{\mathtt{C},\mathtt{avg}}(m,k,d) = O\left(\frac{k^2}{m^2}\log d\right).$$

Remark 2.3.2. *Our formulation assumes that the support size* k *is known. That said, our proposed estimator extends easily to the setting where we only have an upper bound of* k *on the support size, and we seek to output a set of indices containing the support.*

Remark 2.3.3. *We will use the notation* $n^*_{\mathtt{C},\mathtt{avg}}(m,k,d)$ *to denote the sample complexity derived under the assumption of random inputs, and* $n^*_{\mathtt{C},\mathtt{w}}(m,k,d)$ *to denote that the inputs are worst-case or deterministic. The dependence of sample complexity on the problem dimensions* (m,k,d) *remains the same in both cases, but we use this notation to point out the dependence on other parameters in the statement of our results.*

We will present performance guarantees for our estimator in the more general noisy setting in the next section, from which the result above will follow. Our proposed estimator and its analysis applies to a much broader setting involving subgaussian priors. 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 $X_{i,j} \sim \mathrm{subG}(\lambda'_j)$, where λ'_j is the variance parameter for the subgaussian random variable $X_{i,j}$. Our analysis will go through as long as the variance and variance parameters differ only up to a constant factor.

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

For clarity, we restate our assumptions below. These assumptions are required for the analysis of our estimator; the lower bound proof is done under the more restrictive

setting of Gaussian measurement matrix ensemble (which implies a lower bound for the subgaussian ensemble also).

Assumption 2.3.1. *The entries of X_i, $i \in [n]$, are independent and zero-mean with $\mathbb{E}\left[X_{i,j}^2\right] = \lambda_j$ for $\lambda \in \mathcal{S}_{k,d}$ and $X_{i,j} \sim subG(c\lambda_j)$, where c is an absolute constant;*

Assumption 2.3.2. *The $m \times d$ measurement matrices $\Phi_1, \ldots, \Phi_n$ are independent and have entries that are independent and zero-mean with $\mathbb{E}\left[\Phi_i(u,v)^2\right] = 1/m$, $\Phi_i(u,v) \sim subG(c'/m)$, and $\mathbb{E}\left[\Phi_i(u,v)^4\right] = c''/m^2$, where c' and c'' are absolute constants.*

We have restricted λ to binary vectors for ease of presentation. Later, in Section 2.3.1.2, we will show that our sample complexity results extend almost verbatim to a more general setting with the nonzero coordinates of λ taking values between $\lambda_{\min}$ and $\lambda_{\max}$. The only change, in effect, is an additional factor $(\lambda_{\max}/\lambda_{\min})^2$ in the sample complexity of support recovery.

We will work with the more general setting with subgaussian random variables satisfying assumptions 2.3.1 and 2.3.2. In fact, for simplicity, we assume that $X_{i,j}$ and W_i are subgaussian with variance parameter equal to their respective variances, a property known as *strict subgaussianity*. We also note that in Assumption 2.3.2, while subgaussianity will provide an upper bound on the fourth moment, we require the fourth moment to be between c/m^2 and C/m^2 for absolute constants c and C. In essence, we are presenting a unified analysis for Rademacher, Gaussian and other random variables which satisfy these conditions. For notational simplicity, we will fix the value of the constants and take

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

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

Our results also extend to the case when the data vectors are not necessarily sparse in the standard basis for $\mathbb{R}^d$, i.e., the data vectors can be expressed as $X_i = BZ_i$, $i \in [n]$, where B is any known orthonormal basis for $\mathbb{R}^d$ and Z_is have a common support of size k. Under the same generative model as before, but for Z_i this time, Theorem 2.3.1 continues

to hold. This is because when Φ_i has subgaussian entries, the *effective* measurement matrix $\Phi_i B$ also satisfies the properties we mentioned above, namely, it has independent mean zero subgaussian entries with variance and fourth moment of the order $1/m$ and $1/m^2$, respectively.

2.3.1 The closed-form estimator

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

Formally, we compute variance estimates

$$\tilde{\lambda}_i \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} (\Phi_{ji}^\top Y_j)^2, \tag{2.2}$$

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

$$\tilde{\mathcal{S}} = \{(1), ..., (k)\}, \tag{2.3}$$

where (i) denotes the index of the ith largest entry in $\tilde{\lambda}$. This is similar in spirit to the Iterative Hard Thresholding (IHT) algorithm [12] from the compressed sensing literature, where a similar support estimation step followed by least squares is used to estimate the data vectors. The difference is that IHT is an iterative procedure and the least squares step requires $m \geq k$. Note that evaluating $\tilde{\lambda}_i$ requires $O(nm)$ operations, whereby the overall computational complexity of (naively) evaluating our proposed estimator is $O(dnm)$.

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

Lemma 2.3.4. *Let the estimator $\tilde{\lambda}$ be as defined in (2.2). Then, under Assumptions 2.3.1 and 2.3.2 with $c = c' = c'' = 1$, we have that*

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

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

Proof. We can rewrite the estimator in (2.2) as

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

Taking expectation, we note that,

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

where the second step uses the fact that X_j has zero mean entries.

Using our assumption that the columns of Φ_j have independent mean-zero entries with variance $1/m$ and fourth moment $3/m^2$, it follows from Lemma 2.6.13 that

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

which establishes the lemma. □

We work with this biased estimate $\tilde{\lambda}$ and analyze its performance in the next section. Since the bias is the same across all coordinates, it does not affect sorting/thresholding

based procedures. The key observation here is that the expected value of the entries of $\tilde{\lambda}$ for coordinates in the true support exceeds those outside the support, making it an appropriate statistic for support recovery.

We next compute the variance of the estimator, which would give an idea of the performance of the estimator. In particular, it can be used to bound the error probability as a function of the problem parameters by an application of Chebyshev inequality. We will compute the variance for the basic case of Gaussian measurement matrices and Gaussian inputs, with noiseless observations, and it will provide a rough understanding of the sample complexity of the problem. In later sections, we will obtain more precise results by deriving concentration bounds for the estimator. We state our assumptions and the variance bound below.

Assumption 2.3.3. *The entries of X_i, $i \in [n]$, are independent with $X_{i,j} \sim \mathcal{N}(0, \lambda_j)$, $j \in [d]$, and $\lambda \in \mathcal{S}_{k,d}$;*

Assumption 2.3.4. *The $m \times d$ measurement matrices $\Phi_1, \ldots, \Phi_n$ are independent, with entries that are independent and distributed as $\mathcal{N}(0, 1/m)$.*

Lemma 2.3.5. *Let the estimator $\tilde{\lambda}$ be as defined in (2.2). Then, under Assumptions 2.3.3 and 2.3.4, we have in the noiseless setting that*

$$\texttt{var}\left(\lambda_i\right) \leq \begin{cases} c\left(1 + \frac{k^2}{m^2}\right), & \text{if } i \in \mathcal{S}, \\ c' \frac{k^2}{m^2}, & \text{otherwise}, \end{cases}$$

for absolute constants c and c'.

We provide the proof of this lemma in Section 2.6.

We can now use the bound on $\texttt{var}(\tilde{\lambda}_i)$ to obtain a bound on the probability of error per coordinate by an application of Chebyshev inequality. In particular, assuming $k/m > 1$, we get for every $i \in [d]$ and $t > 0$,

$$\Pr\left(\left|\tilde{\lambda}_i - \mathbb{E}\left[\tilde{\lambda}_i\right]\right| > t\right) \leq \frac{k^2}{t^2 m^2 n},$$

which indicates that for constant $t < 1$, the number of samples should scale roughly as $k^2/\delta m^2$ for the probability of error to remain bounded above by δ. This, however, only captures the per- coordinate behaviour of the estimator and an application of the union bound over all coordinates will inevitably lead to a factor of d, indicating that the sample size should scale linearly with d. This is not desirable and can be avoided by using sharper concentration bounds for the estimator. This is the focus of the next section.

2.3.1.1 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 $i \in \mathcal{S}$ and $i' \in \mathcal{S}^c$ cases, then we can distinguish between the two cases. We show that such a separation holds with high probability for subgaussian measurement matrix ensembles satisfying the assumptions in Assumption 2.3.2.

We now present the performance of our estimator.

Theorem 2.3.6. *Let $\tilde{\mathcal{S}}$ be the estimator described in* (2.3)*, and assume that $(\log k)^2 \leq m$ and $k \leq d-1$. Then, under Assumptions 2.3.1 and 2.3.2, $\tilde{\mathcal{S}}$ equals the true support with probability at least $1-\delta$ provided*

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

for an absolute constant c.

Remark 2.3.7. *We note that the result above applies for all k and all $m > (\log k)^2$, and not only to our regime of interest $m < k$. When $\sigma^2 = 0$, $m < k \leq d-1$, and $\delta = 1/3$, we obtain the upper bound claimed in Theorem 2.3.1.*

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

$$\hat{\lambda}_i = \mathbb{1}_{\{\tilde{\lambda}_i \geq \tau\}}. \tag{2.4}$$

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

$$\Pr\left(\tilde{\mathcal{S}} \neq \mathtt{supp}(\lambda)\right) \leq \Pr\left(\hat{\mathcal{S}} \neq \mathtt{supp}(\lambda)\right), \tag{2.5}$$

where $\hat{\mathcal{S}}$ is the support of $\hat{\lambda}$. Using this observation, it suffices to analyze the estimator $\hat{\lambda}$ in (2.4), which will be our focus below.

The proof of Theorem 2.3.6 entails a careful analysis of tails of $\tilde{\lambda}_i$ and uses standard subgaussian and subexponential concentration bounds. To bound the error term in (2.5), we rely on the measurement matrix ensemble satisfying a certain separation condition; we denote this event by $\mathcal{E}$ and describe it in detail shortly. Denoting by $\mathcal{S}$ the support of λ, the error event $\Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}\right)$ can be bounded as

$$\begin{aligned}\Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}\right) &\leq \Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}|\mathcal{E}\right) + \Pr\left(\mathcal{E}^c\right) \\ &\leq \sum_{i\in\mathcal{S}} \Pr\left(\tilde{\lambda}_i < \tau|\mathcal{E}\right) + \sum_{i'\in\mathcal{S}^c} \Pr\left(\tilde{\lambda}_{i'} \geq \tau|E\right) + \Pr\left(\mathcal{E}^c\right). \end{aligned} \tag{2.6}$$

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

Our approach involves deriving tail bounds for $\tilde{\lambda}_i$ conditioned on the measurement matrices, and then choosing a threshold τ to obtain the desired bound for (2.6); we derive lower tail bounds for $i \in \mathcal{S}$ and upper tail bounds for $i' \in \mathcal{S}^c$. The event $\mathcal{E}$ mentioned above corresponds to the measurement ensemble being such that we can find a threshold τ that allows us to separate these bounds.

Specifically, note that

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

where we used $Y_j = \Phi_j X_j + W_j$. Conditioned on Φ_1^n, $\tilde{\lambda}_i$ is a sum of independent subexponential random variables. Using properties of subexponential random variables described in Lemmas 2.6.3 and 2.6.4 in Section 2.6.3, we get that conditioned on the measurement matrices Φ_1^n, the random variable $\tilde{\lambda}_i$ is

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

where c_1 and c_2 are absolute constants and

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

Using standard tail bounds for subexponential random variables given in Lemma 2.6.3 and denoting $\mu_i \stackrel{\text{def}}{=} \mathbb{E}\left[\tilde{\lambda}_i|\Phi_1^n\right] = \frac{1}{n}\sum_{j=1}^n \alpha_{ji}^2, i\in[d]$, we have for $i\in\mathcal{S}$,

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

and for $i'\in\mathcal{S}^c$,

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

We can upper bound the sum of the first two terms in (2.6) by $\delta/2$ by showing that with large probability Φ_1^n takes values for which we get each term above bounded by roughly $\delta' \stackrel{\text{def}}{=} \delta/(4\max\{(d-k),k\})$. In particular, using a manipulation of the expression for exponents, each of the conditional probabilities above will be less than δ' if τ satisfies the following condition for any $i\in\mathcal{S}$ and $i'\in\mathcal{S}^c$:

$$\mu_{i'} + \nu_{i'} \le \tau \le \mu_i - \nu_i, \tag{2.7}$$

where

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

and a similar definition holds for $\nu_{i'}$. Thus, the sufficient condition in (2.7) can be rewritten as

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji}^2 - \frac{1}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq &\max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2 \log\frac{1}{\delta'}\right\} \\ &+ \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji'}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji'}^2 \log\frac{1}{\delta'}\right\}. \qquad (2.8)\end{aligned}$$

Let $\mathcal{E}$ denote the event that for all $i \in \mathcal{S}$ and $i' \in \mathcal{S}^c$, condition (2.8) is satisfied by the measurement matrix ensemble.

We will show that for Φ_1^n drawn from the subgaussian ensemble satisfying assumption 2.3.2, the event $\mathcal{E}$ in fact occurs with high probability. We establish this claim by showing that each term in (2.8) concentrates well around its expected value and roughly $nm^2 \geq ck^2\log(1/\delta')$ suffices to guarantee that the separation required in (2.8) holds with large probability. The following result, which we prove in Section 2.6.1, shows that (2.8) holds with large probability for all pairs $(i, i') \in \mathcal{S} \times \mathcal{S}^c$.

Lemma 2.3.8. *The separation condition* (2.8) *holds simultaneously for all pairs* $(i, i') \in \mathcal{S} \times \mathcal{S}^c$ *with probability at least* $1 - \delta$*, over the choice of* Φ_1^n*,* X_1^n *and* W_1^n*, if* $n \geq c(k/m + \sigma^2)^2 \log(1/\delta')$*, where* $\delta' = \delta/(4\max\{k, d-k\})$*.*

Choosing the probability parameter to be $\delta/2$ in Lemma 2.3.8, we see that the third term in (2.6) can be at most $\delta/2$, leading to an overall error probability of at most δ. Further, noting that $2\log(1/\delta') \geq \log(16k(d-k)/\delta)$, we obtain the result claimed in the theorem. □

Remark 2.3.9. *The separation condition* (2.8) *fails to hold for* $n = 1$*, regardless of which measurement ensemble is used. This is to be expected in our setting of* $m < k$*, since*

from our lower bound for sample complexity stated in Theorem 3.1.1, multiple samples are necessary in the $m < k$ regime.

Remark 2.3.10. *We also note that using the same measurement matrix across samples leads to worse performance for the closed-from estimator. In particular, for the term on the left hand side of* (2.8) *to remain positive (which is necessary since the right hand side is always positive), $m = O(\sqrt{k})$ measurements are required per sample. This dependence arises due to the deviation terms which only decay with m (instead of both m and n when different Φ_i are used). The approximate ML estimator that we will see later can work with the same measurement matrix across samples, but can only handle k/m ratios slightly larger than one (see Figure 2.1 for details). Our lower bound, on the other hand, continues to hold when all Φ_i are equal even with $m = 1$, indicating that better estimators can be designed that can work with $m < \sqrt{k}$ even when the same measurement matrix is used across samples.*

2.3.1.2 Extension to nonbinary variances

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

Assumption 2.3.5. *The entries of X_i, $i \in [n]$, are independent and zero-mean with $\mathbb{E}\left[X_{i,j}^2\right] = \lambda_j$ for $\lambda \in \{u \in \mathbb{R}^d : \|u\|_0 = k, \lambda_{min} \leq u_i \leq \lambda_{max}\}$ and $X_{i,j} \sim subG(c\lambda_j)$, where $0 < \lambda_{min} \leq \lambda_{max}$ and c is an absolute constant. In addition, we assume that $\lambda_{min}/\lambda_{max} > k/(k+m-1)$.*

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

Theorem 2.3.11. *The sample complexity of support recovery under Assumptions 2.3.1 and 2.3.2 satisfies*

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

provided $m \geq (\log k)^2$, where C is an absolute constant.

The techniques used for proving the upper bounds remains essentially the same, and we highlight the key changes in this section.

We start by extending the bias calculation in Lemma 2.3.4 to the more general non-binary setting. We omit the proof since it follows from straightforward calculations.

Lemma 2.3.12. *Let the estimator $\tilde{\lambda}$ be as defined in* (2.2)*. Then, under Assumptions 2.3.1 and 2.3.2 with $c = c' = c'' = 1$, we have that*

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

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

Note that, when λ is binary, $\mathtt{Tr}(K_\lambda) = k$ and the result above reduces to Lemma 2.3.4. WE now provide the proof of Theorem 2.3.11.

Proof. Our final estimate for the support is the same as before, namely, it computes $\tilde{\lambda}$ and declares the indices of the k largest entries as the support. However, as before, we work with a threshold based estimator, with the bias terms in Lemma 2.3.12 being accounted for in the threshold.

Following the same series of arguments as in the binary case, and using the assumption that $\lambda_i \in [\lambda_{\min}, \lambda_{\max}]$, we have that $\Pr\left(\tilde{\mathcal{S}} \neq \mathcal{S}\right) \leq \delta$ can be achieved provided that the following condition holds for every $i \in \mathcal{S}$ and every $i' \in \mathcal{S}^c$:

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

where $\delta' = \delta/(4\max\{k, d-k\})$, and

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

Incorporating the scaling due to $\lambda_{\min}$ and $\lambda_{\max}$ into our concentration bounds in the proof of Lemma 2.3.8, and simplifying, we get that

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

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

□

2.3.2 The maximum likelihood estimator

In this section, we will look at another estimator that can perform support recovery in the $m < k$ regime and is based on the maximum likelihood principle. We will work with the same setting of random inputs as before. In fact, we will restrict to a Gaussian prior on the inputs and use the specific form of the output covariance matrix that results due to the linear model to develop the estimator. We will use the variance parameter λ of the prior to capture the common support structure of the inputs. Our approach at a high level is to express the sample covariance matrix as a perturbed version of the population covariance matrix that is parameterized by the measurement matrices and the prior parameter λ. We then use a Gaussian approximation on the distribution of the perturbation term, and set up a maximum likelihood problem to estimate the prior parameter. More specifically, we find the approximate maximum likelihood estimate of λ using a modified reweighted minimization procedure. Empirically, the proposed algorithm succeeds in exactly recovering the common support with high probability in the $m \geq k$ regime with n of the order of m and in the $m < k$ regime with larger n.

To capture the latent structure in X_i, we assume that $X_i \sim \mathcal{N}(0, K_\lambda)$ where recall from the previous section that $K_\lambda = \texttt{diag}\,(\lambda)$. This multivariate Gaussian prior to model sparsity was first introduced in [87]. Also, unlike the previous closed-form estimator, the maximum likelihood estimator that we will see in this section can work in the setting where the same measurement matrix is used across samples. We will derive the estimator

for this setting, but the same approach can be used in the case where different measurement matrices are used across samples. The observations Y_i are therefore distributed as $\mathcal{N}(0, \Phi K_\lambda \Phi^\top + \sigma^2 I)$ and the goal is to estimate the common support $\mathcal{S}$ from $\{Y_i, \Phi_i\}_{i=1}^n$. We observe that under the prior model above, $\mathcal{S} = \texttt{supp}\,(X_i) = \texttt{supp}\,(\lambda)$, since $\lambda_j = 0$ if and only if $X_{ij} = 0$ almost surely. Hence, support recovery from multiple samples is equivalent to recovering the support of λ.

2.3.2.1 Gaussian approximation based support recovery

Let $K \in \mathbb{R}^{m \times m}$ denote the covariance matrix of the observations. Then, in the noiseless case, we have $K = \Phi K_\lambda \Phi^\top$, which can be rewritten after vectorizing as $\texttt{vec}\,(K) = (\Phi \odot \Phi)\lambda$, where $\odot$ denotes the Khatri-Rao product [38]. The support recovery problem is to then recover the support of the sparse non-negative vector λ from K. However, instead of solving the problem over a discrete variable (the support), we will solve for the variance parameter itself and show that this can be done efficiently using a nonnegative quadratic program. We note that reformulating support recovery as a variance estimation problem (using a convex optimization procedure) was first considered in [54] as

$$\min_{\lambda} \|\lambda\|_1 \text{ s.t. } (\Phi \odot \Phi)\lambda = \texttt{vec}\,(K)\,. \tag{2.9}$$

This model is analyzed in [53], and conditions under which the model is identifiable are derived. If we had access to the true covariance matrix K (which corresponds to the $n \to \infty$ case), then we could work with the system of equations $K = \Phi K_\lambda \Phi^\top$ to recover the support of λ which, in turn, would give us the common support of X_is. For finite n, we can use the sample covariance matrix $\hat{K} = (1/n)\sum_{i=1}^n Y_i Y_i^\top$ as an estimate for K, but we need to account for the error arising due to finite samples. In this section, we derive the statistics of the error due to finite sample approximation to K, and then find the ML estimate of λ. More precisely, the sample covariance matrix can be written as a noisy version of the true covariance matrix as

$$\hat{K} = K + K_{\text{err}}, \tag{2.10}$$

where K_{err} represents the error matrix. Equivalently, vectorizing the matrices on either side of (2.10), we get

$$R = (\Phi \odot \Phi)\lambda + N, \tag{2.11}$$

where $R \stackrel{\text{def}}{=} \mathtt{vec}\left(\hat{K}\right)$ and $N \stackrel{\text{def}}{=} \mathtt{vec}\left(K_{\text{err}}\right)$. We now proceed to find the approximate ML estimate of λ. To that end, we first derive the statistics of the error N.

Our starting point is the following lemma, which provides the mean and covariance matrix of the vectorized error N. Also, for a random vector X, we will use the shorthand $\mathtt{cov}\left(X\right)$ to denote its covariance matrix $\mathbb{E}\left[(X - \mathbb{E}\left[X\right])(X - \mathbb{E}\left[X\right])^\top\right]$.

Lemma 2.3.13. *Consider $\{Y_i\}_{i=1}^n$ drawn iid from $\mathcal{N}(0, K)$. Let $\hat{K}$ denote the sample covariance formed using $\{Y_i\}_{i=1}^n$ and let $N = \mathtt{vec}\left(\hat{K} - K\right)$. Further, let $B = \mathtt{cov}\left(\mathtt{vec}\left(ZZ^\top\right)\right)$ where $Z \sim \mathcal{N}(0, I)$ and let C be a matrix satisfying $K = CC^\top$. Then,*

$$\mathbb{E}\left[N\right] = 0, \text{ and } \mathtt{cov}\left(N\right) = \frac{1}{n}(C \otimes C)B(C \otimes C)^\top.$$

Proof. The mean computation is straightforward:

$$\mathbb{E}\left[N\right] = \frac{1}{n}\sum_{i=1}^{n} \mathtt{vec}\left(\mathbb{E}\left[Y_i Y_i^\top - K\right]\right) = 0.$$

The covariance matrix can be computed as follows:

$$\begin{aligned}\mathtt{cov}\left(N\right) &= \mathtt{cov}\left(\mathtt{vec}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^\top - K\right)\right) \\ &= \frac{1}{n}\mathtt{cov}\left(\mathtt{vec}\left(Y_1 Y_1^\top - K\right)\right) = \frac{1}{n}\mathtt{cov}\left(\mathtt{vec}\left(YY^\top\right)\right),\end{aligned}$$

where we used the fact that $Y_1, \ldots, Y_n$ are independent and identically distributed. We now represent Y as $Y = CZ$, where $Z \sim \mathcal{N}(0, I)$ and C is a matrix such that $K = CC^\top$. Using this, we obtain

$$\begin{aligned}\mathtt{cov}\left(N\right) &= \frac{1}{n}\mathtt{cov}\left(\mathtt{vec}\left(CZZ^\top C^\top\right)\right) \\ &= \frac{1}{n}\mathtt{cov}\left((C \otimes C)\mathtt{vec}\left(ZZ^\top\right)\right)\end{aligned}$$

$$= \frac{1}{n}(C \otimes C)B(C \otimes C)^\top,$$

where $B \stackrel{\text{def}}{=} \texttt{cov}\left(\texttt{vec}\left(ZZ^\top\right)\right)$. □

For our model, $K = \Phi K_\lambda \Phi^\top + \sigma^2 I$. Letting $C = \Phi D^{\frac{1}{2}}$, with $D = K_\lambda + \sigma^2 \Phi^\dagger \Phi^{\top\dagger}$, and using the lemma above, we get

$$\begin{aligned}\texttt{cov}\,(N) &= \frac{1}{n}(\Phi D^{\frac{1}{2}} \otimes \Phi D^{\frac{1}{2}})B(\Phi D^{\frac{1}{2}} \otimes \Phi D^{\frac{1}{2}}) \\ &= \frac{1}{n}(\Phi \otimes \Phi)(D^{\frac{1}{2}} \otimes D^{\frac{1}{2}})B(D^{\frac{1}{2}} \otimes D^{\frac{1}{2}})(\Phi \otimes \Phi)^\top, \end{aligned} \tag{2.12}$$

where the second step uses the property that $UV \otimes XY = (U \otimes X)(V \otimes Y)$. The $d^2 \times d^2$ covariance matrix B of $\texttt{vec}\left(ZZ^\top\right)$ can be computed explicitly for a given d and it can be verified that the entries of B lie in $\{0, 1, 2\}$.

We give an example for the case when $d = 3$.

Example The matrix B for $N = 3$.

Let $Z = [Z_1, Z_2, Z_3]^\top$ with $Z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, $i \in \{1, 2, 3\}$. We wish to find the covariance matrix of $\texttt{vec}\left(ZZ^\top\right)$. For example, the $(1,1)^{th}$and $(1,2)^{th}$ entries can be computed as follows:

$$\begin{aligned} B_{1,1} &= \mathbb{E}\left[Z_1^4\right] - \left(\mathbb{E}\left[Z_1^2\right]\right)^2 = 3 - 1 = 2 \\ B_{1,2} &= \mathbb{E}\left[Z_1^3 Z_2\right] - \mathbb{E}\left[Z_1^2\right]\mathbb{E}\left[Z_1 Z_2\right] = 0. \end{aligned}$$

Computing the remaining entries in a similar way, we get

$$B = \mathrm{cov}(\mathrm{vec}(ZZ^\top)) = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}.$$

We also note that B is rank-deficient with $\texttt{rank}(B) = d(d+1)/2$.

For the noiseless case, we have $D = K_\lambda$ and we can further simplify (2.12) by exploiting the structure of B. Specifically, it can be shown that B can be expressed as $I_{m^2} + Q$, where Q denotes a permutation matrix and I_{m^2} denotes the $m^2 \times m^2$ identity matrix. Using this fact and the structure of $K_\lambda^{\frac{1}{2}} \otimes K_\lambda^{\frac{1}{2}}$, we get

$$\begin{aligned} M &\stackrel{\text{def}}{=} \texttt{cov}\,(N) \\ &= \frac{1}{n}(\Phi \otimes \Phi)(K_\lambda^{\frac{1}{2}} \otimes K_\lambda^{\frac{1}{2}})(I_{m^2} + Q)(K_\lambda^{\frac{1}{2}} \otimes K_\lambda^{\frac{1}{2}})(\Phi \otimes \Phi)^\top \\ &= \frac{1}{n}(\Phi \otimes \Phi)B(K_\lambda \otimes K_\lambda)(\Phi \otimes \Phi)^\top. \end{aligned} \tag{2.13}$$

In the next section, we use these statistics to derive an approximate ML estimate of λ.

2.3.2.2 Maximum Likelihood Estimation of λ

We consider the model derived in the previous section:

$$R = A_{\texttt{KR}}\lambda + N, \tag{2.14}$$

where $A_{\texttt{KR}} \stackrel{\text{def}}{=} (\Phi \odot \Phi)$. We seek the ML estimate of λ from R. It is important to note that the statistics of the noise N also depends on λ.

Since R, $A_{\mathtt{KR}}\lambda$ and N are vectorized versions of $m \times m$ symmetric matrices, they lie in an $\frac{m(m+1)}{2}$ dimensional subspace of $\mathbb{R}^{m^2}$. We therefore restrict our attention to the $\frac{m(m+1)}{2}$ linearly independent equations in (2.14). This can be done by pre-multiplying (2.14) by a projection matrix $P \in \mathbb{R}^{\frac{m(m+1)}{2} \times m^2}$, formed using a subset of the rows of I_{m^2} that picks the $\frac{m(m+1)}{2}$ independent entries. Thus,

$$R_P = A_{\mathtt{KR},P}\lambda + N_P,$$

where $R_P \stackrel{\text{def}}{=} PR$, $A_{\mathtt{KR},P} \stackrel{\text{def}}{=} PA_{\mathtt{KR}}$, and $N_P \stackrel{\text{def}}{=} PN$. Further, we approximate the distribution of N_P by $\mathcal{N}(0, M_P)$, where $M_P = PMP^\top$ and M is the noise covariance matrix derived in the previous section. This Gaussian approximation is motivated from the fact that the noise vector N is a sum of i.i.d. random vectors, i.e.,

$$N = \frac{1}{n}\left(\sum_{i=1}^{n} \mathtt{vec}\left(Y_i Y_i^\top - \mathbb{E}\left[Y_i Y_i^\top\right]\right)\right) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} U_i,$$

which implies, from the central limit theorem, that as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} U_i \stackrel{d.}{\to} \mathcal{N}(0, M)$.

Using this, the approximate ML estimate of λ, which we denote λ_{ML}, can be found by solving the following optimization problem:

$$\lambda_{\text{ML}} = \arg\max_{\lambda \geq 0} \; p(r_P; \lambda), \tag{2.15}$$

where

$$p(r_P; \lambda) = \frac{1}{(2\pi)^{\frac{m(m+1)}{4}} |M_P|^{\frac{1}{2}}} \exp\left(-\frac{(r_P - A_{\mathtt{KR},P}\lambda)^\top M_P^{-1}(r_P - A_{\mathtt{KR},P}\lambda)}{2}\right),$$

and r_P denotes an instantiation of R_P. Simplifying (2.15), we get

$$\lambda_{\text{ML}} = \arg\min_{\lambda \geq 0} \; \log|M_P| + (r_P - A_{\mathtt{KR},P}\lambda)^\top M_P^{-1}(r_P - A_{\mathtt{KR},P}\lambda). \tag{2.16}$$

The objective function in (2.16) is nonconvex in λ since M_P also depends on λ, and

Algorithm 1: Modified Reweighted NNQP (MRNNQP)

Input: Measurement matrix Φ, vectorized version r_P of upper triangular entries of sample covariance matrix of $\{Y_i\}_{i=1}^n$, initial value $\lambda^{(0)} = (1, \ldots, 1)^\top$, $K_\lambda^{(0)} = \mathrm{diag}(\lambda^{(0)})$, $i = 1$

While (not converged) **do**

$M_P^{(i)} \leftarrow \frac{1}{n} P(\Phi \otimes \Phi) B (K_\lambda^{(i-1)} \otimes K_\lambda^{(i-1)})(\Phi \otimes \Phi)^\top P^\top$

$b^{(i)} = -A_{\mathtt{KR},P}^\top {M_P^{(i)}}^{-1} r_P$

$Q^{(i)} = A_{\mathtt{KR},P}^\top {M_P^{(i)}}^{-1} A_{\mathtt{KR},P}$

$\lambda^{(i)} = \mathrm{NNQP}(Q^{(i)}, b^{(i)})$

$K_\lambda^{(i)} = \mathrm{diag}(\lambda^{(i)})$

$i = i + 1$

end While

Output: Support of $\lambda^{(i)}$

is difficult to optimize directly. In the next section, we propose a heuristic technique to solve the optimization problem.

2.3.2.3 Modified Reweighted Minimization

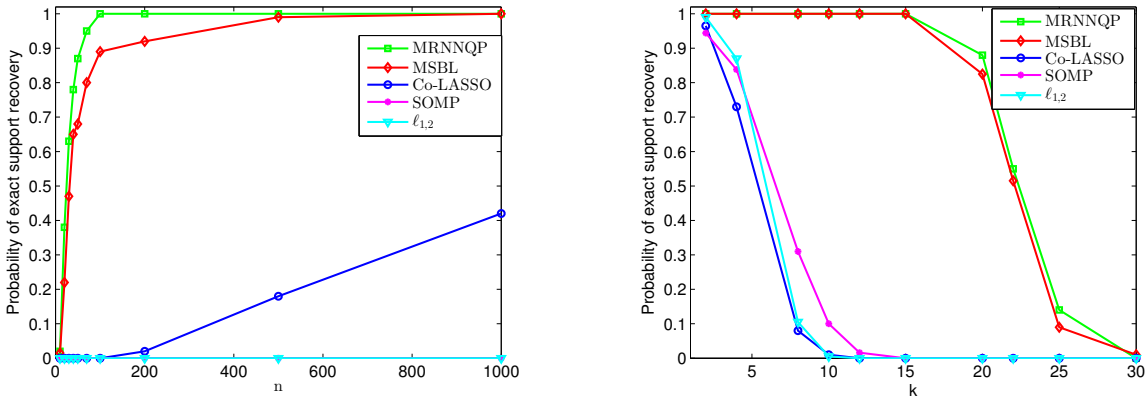
In this section, we propose a modified reweighted minimization approach to solve (2.16). We fix M_P, solve the resulting convex non-negative quadratic problem, re-compute M_P using the new λ, and iterate.

Now, to solve the convex non-negative quadratic program (NNQP)

$$\arg\min_{\lambda \geq 0} \ (r_P - A_{\mathtt{KR},P}\lambda)^\top M_P^{-1} (r_P - A_{\mathtt{KR},P}\lambda),$$

we use the iterative technique of [70], which gives the following entry-wise update for λ in the $(i+1)$th iteration:

$$\lambda_j^{(i+1)} = \lambda_j^{(i)} \left(\frac{-b_j + \sqrt{b_j^2 + 4(Q^+\lambda^{(i)})_j (Q^-\lambda^{(i)})_j}}{2(Q^+\lambda^{(i)})_j} \right),$$



(a) d=40, m=20, k=25 (b) d=70, m=20, n=50

Figure 2.1: Support recovery performance of different algorithms.

where $b = -A_{\mathtt{KR},P}^{\top} M_P^{-1} r_P$, $Q = A_{\mathtt{KR},P}^{\top} M_P^{-1} A_{\mathtt{KR},P}$, $Q^+ = \max(Q, 0)$, $Q^- = \max(-Q, 0)$, with $\max(Q, 0)$ representing the entry-wise maximum of the elements of Q and 0.

Thus, our approach is as follows: we approximate the noise covariance M_P by its zeroth-order Taylor expansion around a previous estimate of λ and then minimize the resulting cost function over λ keeping M_P fixed. This can be viewed as an iterative reweighted minimization [16] technique where we only consider the zeroth order term in the Taylor expansion, since gradient computation is difficult. The steps are summarized in Algorithm 1. The computational complexity of the algorithm is dominated by the computation of M_P, which requires $O(m^4 d^2)$ operations. However, increasing the number of samples only affects the computation of r (which can be computed in $O(n)$ operations).

We point out some important aspects of the model (2.11) and differences between our algorithm and existing algorithms in the literature. The statistics of the error term N depends on n as well as on the parameter λ that has to be estimated, as can be seen from (2.13). As a result of this parameter-dependent noise, the maximum likelihood cost function is nonconvex in λ and difficult to optimize. The Co-LASSO algorithm [54], which also uses the sample covariance matrix to estimate λ, does not account for the statistics of the noise/error arising because of the difference between the true covariance and its finite sample based estimate. Therefore, the algorithm performs well only when n is large,

i.e., when the error term is negligible. As we illustrate in the next section, the proposed algorithm performs well at a much smaller n. Also, under our generative model for the inputs, the $\ell_{1,2}$ penalty algorithm [44] and simultaneous OMP [78] perform poorly in the $m < k$ regime.

Another interesting feature of our algorithm is that the key step, namely, Step 6 in Algorithm 1, involves solving a nonnegative quadratic program. In particular, no sparsity-promoting penalty is required. Similar observations were made in [29], where the authors note that a non negative least squares program can be used for recovering nonnegative sparse vectors without explicit sparsity-inducing regularization.

2.3.2.4 Simulation Results

In this section, we study the support recovery performance of the proposed algorithm through simulations.

For a given set of (m, k, d, n) values, we generate the following: an $m \times d$ measurement matrix Φ with entries $\Phi_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/m)$, a support $\mathcal{S} \subset [d]$ with $|\mathcal{S}| = k$ chosen uniformly at random from $\binom{d}{k}$ possibilities, $\lambda \in \{0, 1\}^d$ with support $\mathcal{S}$, $\{X_i\}_{i=1}^n$ drawn independently from $\mathcal{N}(0, \texttt{diag}\,(\lambda))$. For each trial, The algorithm is provided with Φ and $\{Y_i\}_{i=1}^n$ generated according to the linear model. We run the algorithm 200 times, and a trial is declared successful if the algorithm exactly recovers the true support. The objective value decreases as the iterations proceed and stabilizes after about 20 iterations.

Figure 2.1 shows the probability of successful recovery of the proposed algorithm, the Co-LASSO approach from [54], the M-SBL algorithm [88], simultaneous OMP (SOMP) [78], and the $\ell_{1,2}$ penalty algorithm [44], as a function of n and k, respectively. Both the proposed algorithm and M-SBL, which use a maximum likelihood based approach to estimate λ show similar performance, with the proposed algorithm performing slightly better in the low n regime. The Co-LASSO approach requires much larger n for reliable support recovery, while SOMP and $\ell_{1,2}$ minimization perform well only in the $m > k$ regime. Thus, our proposed algorithm provides competitive performance with the attractive benefit that its complexity scales linearly with n.

2.4 The case of deterministic inputs

In this section, we analyze the closed form estimator for *deterministic, worst case* inputs and derive an upper bound on the sample complexity that matches with the result in Theorem 2.3.6 with respect to dependence on the problem dimensions. In particular, let vectors $x_1, \ldots, x_n$ in $\mathbb{R}^d$ have a common support $\mathcal{S} \subset [d]$ of cardinality k. For each of these vectors, we have access to noisy linear measurements of the form $Y_i = \Phi_i x_i + W_i$, $i \in [n]$. Here, $\Phi_i \in \mathbb{R}^{m \times d}$ with $m < d$ are called the measurement matrices and $W_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$ is noise. The goal is to recover the support $\mathcal{S}$ using $\{Y_i, \Phi_i\}_{i=1}^n$. An estimator for $\mathcal{S}$ is a mapping $\hat{\mathcal{S}} : \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times d \times n} \to \binom{[d]}{k}$, where $\binom{[d]}{k}$ denotes the set of all subsets of $[d]$ of cardinality k. We assume that the estimator has knowledge of k and consider the probability of exact recovery, $\Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}\right)$, as our recovery criterion. We note that one could also consider the setting where $|\mathcal{S}| \leq k$. The estimator that we consider here would output an $\hat{\mathcal{S}}$ that *contains* the true support with high probability. In this work, however, we assume that the true support has cardinality exactly k.

We will assume that the measurement matrices satisfy Assumption 2.3.4. For the inputs, we make the following assumption.

Assumption 2.4.1. *The d-dimensional inputs $x_1, \ldots, x_n$ are such that $\texttt{supp}(x_i) = \mathcal{S}$, for all $i \in [n]$, where $\mathcal{S} \subset [d]$ is a fixed set of cardinality k. Further, $|x_{iu}| \in [x_{\min}, x_{\max}]$, for all $i \in [n]$, $u \in \mathcal{S}$, where $x_{\min}, x_{\max} \in \mathbb{R}$.*

As before, our focus will be on the measurement-constrained setting where we obtain only $m < k$ measurements per sample, although we will provide results for both $m \geq k$ and $m < k$. We now define the fundamental quantity of interest for us.

Definition 2.4.1. *For $m, k, d \in \mathbb{N}$, the sample complexity of common support recovery $n^*_{\texttt{C},\texttt{w}}(m, k, d)$ is the minimum number of samples n for which we can find an estimator that satisfies*

$$\Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}\right) \leq \delta, \quad \forall \mathcal{S} \in \binom{[d]}{k}. \tag{2.17}$$

Our main result is the following.

Theorem 2.4.1. *The sample complexity of common support recovery under Assumptions 2.3.4 and 2.4.1, for $m \geq 2\log(d/\delta)$ satisfies*

$$n^*_{\mathtt{C},\mathtt{w}}(m,k,d) = O\left(\frac{x_{\max}^4}{x_{\min}^4}\max\left\{\left(\frac{k}{m}+\frac{\sigma^2}{x_{\max}^2}\right)\log\frac{d}{\delta}, \left(\frac{k}{m}+\frac{\sigma^2}{x_{\max}^2}\right)^2\log\frac{d}{\delta}\right\}\right).$$

As a special case, in the noiseless setting with $m < k$ and for constant error probability, we have the following corollary.

Corollary 2.4.2. *In the noiseless setting, with $m \geq 2\log 3d$, $m < k \leq d-1$, and $\delta = 1/3$, we have,*

$$n^*_{\mathtt{C},\mathtt{w}}(m,k,d) = O\left(\frac{x_{\max}^4}{x_{\min}^4}\frac{k^2}{m^2}\log d\right).$$

We provide the proof of Theorem 2.4.1 in the next section.

2.4.1 Analysis of the closed-form estimator

We will analyze the closed form estimator from Section 2.3.1, but instead of random inputs, here we will consider deterministic inputs $x_1, \ldots, x_n$. To see why the analysis from the random input case does not extend in a straightforward way to this case, we first recall the form of the estimator. Let $\Phi_{iu} \in \mathbb{R}^m$ denote the uth column of Φ_i. We first compute proxy samples $\hat{X}_1, \ldots, \hat{X}_n$ with entries

$$\hat{X}_{iu} \stackrel{\text{def}}{=} \Phi_{iu}^\top Y_i = \Phi_{iu}^\top \Phi_i x_i + \Phi_{iu}^\top W_i, \quad u \in [d], \tag{2.18}$$

and then the compute sample variance along each coordinate as

$$\tilde{\lambda}_u \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \hat{X}_{iu}^2, \quad u \in [d]. \tag{2.19}$$

The support estimate $\tilde{\mathcal{S}}$ consists of the k indices of $\tilde{\lambda}$ with the largest value. Analyzing the estimator would basically involve obtaining tail bounds for the random variable above. Considering the noiseless case first, note that each summand in (2.19) is of the form

$(\Phi_{iu}^\top \Phi_i x_i)^2$, and can be viewed as a quadratic in either x_i or $\Phi_{iu}^\top \Phi_i$.

When x_is are random and subgaussian with independent coordinates, we can exploit the quadratic form in x_i to obtain a tail bound using standard subexponential concentration (after conditioning on Φ_i) as we did in the proof of Theorem 2.3.6. On the other hand, when x_i are deterministic, the summands in (2.19) are quadratic in $\Phi_{iu}^\top \Phi_i$, resulting in a heavy-tailed random variable, and standard methods based on bounding the MGF do not work.

We explain in the next section how a careful analysis involving conditioning on a certain *column* of Φ_i followed by a moment based bound can be used to get exponential tail bounds for heavy-tailed random variables. The analysis in Section 2.3 also deals with heavy-tailed random variables that are functions of Φ_i, but uses a more elementary approach which would not work here.

2.4.1.1 A separation condition for support recovery

We will analyze the error probability of the threshold-based version of the closed-form estimator as before. In particular, we will use the estimate $\hat{\lambda} \stackrel{\text{def}}{=} \mathbb{1}_{\{\tilde{\lambda} \geq \tau\}}$, for an appropriate threshold τ, since $\Pr\left(\tilde{\mathcal{S}} \neq \mathcal{S}\right) \leq \Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}\right)$, where $\tilde{\mathcal{S}}$ and $\hat{\mathcal{S}}$ denote, respectively, the supports of $\tilde{\lambda}$ and $\hat{\lambda}$. The error probability $\Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}\right)$ will essentially be determined by the tail behaviour of the variance estimate $\tilde{\lambda}$. Recall from the last section that variance estimate is an average of random variables of the form $(\Phi_{iu}^\top \Phi_i x_i + \Phi_{iu}^\top W_i)^2$. The $\Phi_{iu}^\top \Phi_i x_i$ term will be indicative of whether the coordinate u lies in the support or not, since it will have a $\|\Phi_{iu}\|_2^2$ term only when $u \in \mathcal{S}$.

The analysis is greatly simplified once we condition on Φ_{iu}, because then the summands in (2.19) are noncentral chi-square distributed, for which tail bounds can be obtained using standard methods. The error probability can be made small provided these tail probabilities (parameterized by Φ_{iu}) can be made small, which eventually leads to a condition on the measurement ensemble. We will show using tail bounds for heavy-tailed random variables that this condition is satisfied with high probability for the Gaussian ensemble when the parameters (n, m, k, d) scale as indicated in Theorem 2.4.1, thus finishing the

proof.

The probability of error can be bounded as

$$\Pr\left(\hat{\mathcal{S}} \neq \mathcal{S}\right) \leq \sum_{u \in \mathcal{S}} \Pr\left(\tilde{\lambda}_u < \tau | \mathcal{E}\right) + \sum_{u' \in \mathcal{S}^c} \Pr\left(\tilde{\lambda}_{u'} \geq \tau | \mathcal{E}\right) + \Pr\left(\mathcal{E}^c\right), \tag{2.20}$$

where $\mathcal{E}$ denotes the event that the measurement ensemble satisfies a certain condition, which we will describe shortly. For the right hand side to remain below δ, we require the summands in the first two terms to be at most $\delta/(3\max\{k, d-k\})$. For simplicity, we will work with a requirement of $\delta/3d$. Now, using (2.18) and (2.19), we can see that $\hat{X}_{iu}|\Phi_{iu} \sim \mathcal{N}(\mu_i, \nu_i^2)$ for $u \in \mathcal{S}$ with

$$\mu_i = \|\Phi_{iu}\|_2^2 x_{iu},$$

and

$$\nu_i^2 = \frac{\|\Phi_{iu}\|_2^2}{m} \sum_{v \in \mathcal{S} \setminus \{u\}} x_{iv}^2 + \sigma^2 \|\Phi_{iu}\|_2^2.$$

Similarly, we have $\hat{X}_{iu'}|\Phi_{iu'} \sim \mathcal{N}(0, \nu_i'^2)$, for $u' \in \mathcal{S}^c$, where

$$\nu_i'^2 = \frac{\|\Phi_{iu'}\|_2^2}{m} \sum_{v \in \mathcal{S}} x_{iv}^2 + \sigma^2 \|\Phi_{iu'}\|_2^2.$$

A direct application of Lemma 2.6.11 then yields, for every $u \in \mathcal{S}$,

$$\Pr\left(\tilde{\lambda}_u < \tau | \{\Phi_{iu}\}_{i=1}^n\right) \leq \exp\left(\frac{-n^2(\mu - \tau)^2}{4(\sum_{i=1}^n \nu_i^4 + \nu_i^2 \mu_i^2)}\right),$$

where $\mu \stackrel{\text{def}}{=} \mathbb{E}\left[\tilde{\lambda}_u | \{\Phi_{iu}\}_{i=1}^n\right]$. For $u' \in \mathcal{S}^c$, we can obtain in a similar manner from Lemma 2.6.11,

$$\Pr\left(\tilde{\lambda}_{u'} \geq \tau | \{\Phi_{iu'}\}_{i=1}^n\right) \leq \exp\left(-\min\left\{\frac{n^2(\tau - \mu')^2}{16\sum_{i=1}^n \nu_i'^4}, \frac{n(\tau - \mu')}{8\max_{i \in [n]} \nu_i'^2}\right\}\right),$$

where $\mu' \stackrel{\text{def}}{=} \mathbb{E}\left[\tilde{\lambda}_{u'} | \{\Phi_{iu}\}_{i=1}^n\right]$. For the missed detection and false alarm probabilities above to remain bounded above by $\delta/3d$, we require

$$\tau \leq \mu - \sqrt{\frac{4}{n^2}\sum_{i=1}^n (\nu_i^4 + \mu_i^2\nu_i^2)\log\frac{3d}{\delta}},$$

and

$$\tau \geq \mu' + \max\left\{\sqrt{\frac{16}{n^2}\sum_{i=1}^n \nu_i'^4 \log\frac{3d}{\delta}}, \frac{8}{n}\max_{i\in[n]}\nu_i'^2\log\frac{3d}{\delta}\right\}.$$

Therefore, for the existence of a threshold τ, we can see upon simplification that it suffices to have

$$\mu - \mu' > \sqrt{\frac{4}{n^2}\sum_{i=1}^n (\nu_i^4 + \nu_i^2\mu_i^2)\log\frac{3d}{\delta}} + \max\left\{\sqrt{\frac{16}{n^2}\sum_{i=1}^n \nu_i'^4 \log\frac{3d}{\delta}}, \frac{8}{n}\max_{i\in[n]}\nu_i'^2\log\frac{3d}{\delta}\right\}. \tag{2.21}$$

A simple calculation shows that the conditional mean of the estimator under the $u \in \mathcal{S}$ and $u' \in \mathcal{S}^c$ cases are separated roughly by a constant term (after averaging over the measurement matrices), which makes the distinction between the two cases possible. In particular,

$$\mu = \frac{1}{n}\sum_{i=1}^n \left(x_{iu}^2\|\Phi_{iu}\|_2^4 + \|\Phi_{iu}\|_2^2\left(\frac{1}{m}\sum_{v\in\mathcal{S}\setminus\{u\}} x_{iv}^2 + \sigma^2\right)\right),$$

and

$$\mu' = \frac{1}{n}\sum_{i=1}^n \|\Phi_{iu}\|_2^2\left(\frac{1}{m}\sum_{v\in\mathcal{S}} x_{iv}^2 + \sigma^2\right).$$

Substituting this into (2.21) and simplifying, we can rewrite the condition as

$$\frac{x_{\min}^2}{x_{\max}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\|\Phi_{iu}\|_2^4-\frac{1}{m}\|\Phi_{iu}\|_2^2\right) > \sqrt{\frac{4}{n^2}\left(\frac{k-1}{m}+\frac{\sigma^2}{x_{\max}^2}\right)^2\sum_{i=1}^{n}\|\Phi_{iu}\|_2^4\log\frac{3d}{\delta}}$$
$$+\sqrt{\frac{4}{n^2}\left(\frac{k-1}{m}+\frac{\sigma^2}{x_{\max}^2}\right)\sum_{i=1}^{n}\|\Phi_{iu}\|_2^6\log\frac{3d}{\delta}}+\sqrt{\frac{16}{n^2}\left(\frac{k}{m}+\frac{\sigma^2}{x_{\max}^2}\right)^2\sum_{i=1}^{n}\|\Phi_{iu'}\|_2^4\log\frac{3d}{\delta}}$$
$$+\frac{8}{n}\left(\frac{k}{m}+\frac{\sigma^2}{x_{\max}^2}\right)\max_{i\in[n]}\|\Phi_{iu'}\|_2^2\log\frac{3d}{\delta}, \tag{2.22}$$

for every $(u,u')\in\mathcal{S}\times\mathcal{S}^c$.

2.4.1.2 Separation condition for the Gaussian ensemble

We will show that when the measurement ensemble is Gaussian as described in Assumption 2.3.4, the separation condition in (2.22) is satisfied with high probability for a certain regime of the parameters (n,m,k,d). We will derive upper and lower bounds on the right hand side and left hand side respectively in (2.22), that hold with high probability, which after simplification will finally result in a condition on the parameters as stated in Theorem 2.4.1. Note that this translates to obtaining tail bounds for the random variable $(1/n)\sum_{i=1}^{n}\|\Phi_{iu}\|_2^{2q}$ with $q=2,3$. It is easy to see that $\|\Phi_{iu}\|_2^2$ is chi-square distributed (after scaling by m), and $\|\Phi\|_2^{2q}$ is therefore a heavy-tailed random variable, and so MGF based methods cannot be used here. We will see that a bound on the moments can be used to get exponential tail bounds, even when the MGF is unbounded.

The proofs for results in this section can be found in Section 2.6.2.

We will fix $q=3$ and derive our results; the same arguments can be used for the $q=2$ case as well. Define $Z\stackrel{\text{def}}{=}|(1/n)\sum_{i=1}^{n}(\|\Phi_{iu}\|_2^6-\mathbb{E}\left[\|\Phi_{iu}\|_2^6\right]|$ and note that for all $p\geq 1$,

$$\Pr\left(Z\geq e(\mathbb{E}\left[Z^p\right])^{\frac{1}{p}}\right)=\Pr\left(Z^p\geq e^p\mathbb{E}\left[Z^p\right]\right)\leq e^{-p}. \tag{2.23}$$

Further, for all $p\geq 2$, if we can show that $(\mathbb{E}\left[Z^p\right])^{\frac{1}{p}}\leq cp^{\beta}$ for some $\beta>0$, then together with the previous inequality it implies that $\Pr\left(Z\geq ecp^{\beta}\right)\leq e^{-p}$, or, equivalently, for

$t > 0$, that

$$\Pr\left(Z \geq t\right) \leq \exp(-(t/ec)^{\frac{1}{\beta}}). \tag{2.24}$$

We now need to determine an upper bound on $\|Z\|_{\mathcal{L}_p} \stackrel{\text{def}}{=} (\mathbb{E}\left[Z^p\right])^{\frac{1}{p}}$. We show such a moment bound, resulting in the following lemma.

Lemma 2.4.3. *For every $t > 0$, there exists an absolute constant C such that*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\|\Phi_{iu}\|_2^6 - \mathbb{E}\left[\|\Phi_{iu}\|_2^6\right])\right| \geq t\right) \leq \exp\left(-C\min\left\{nt, (m^3nt)^{\frac{1}{4}}, nt^2\right\}\right).$$

A similar result can be obtained for the $(1/n)\sum_{i=1}^{n}\|\Phi_{iu}\|_2^4$ term in (2.22) using the same technique, and we omit the proof for this result.

Lemma 2.4.4. *For every $t > 0$, there exists an absolute constant C such that*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\|\Phi_{iu}\|_2^4 - \mathbb{E}\left[\|\Phi_{iu}\|_2^4\right])\right| \geq t\right) \leq \exp\left(-C\min\left\{nt, (m^2nt)^{\frac{1}{3}}, nt^2\right\}\right).$$

Together with the fact that $\mathbb{E}\left[\|\Phi_{iu}\|_2^4\right] = 1 + 2/m$ and $\mathbb{E}\left[\|\Phi_{iu}\|_2^6\right] = 1 + 6/m + 8/m^2$, the results above give upper and lower bounds that hold with high probability on all but the $\max_{i\in[n]}\|\Phi_{iu'}\|_2^2$ term in (2.22). The latter can be bounded with high probability using concentration for chi-squared random variables and a union bounding step, as given by the following lemma.

Lemma 2.4.5. *Let $\mu_{\max} \stackrel{\text{def}}{=} \mathbb{E}\left[\max_{i\in[n]}\|\Phi_{iu}\|_2^2\right]$. Then, for every $t > 0$,*

$$\Pr\left(\max_{i\in[n]}\|\Phi_{iu}\|_2^2 \geq \mu_{\max} + t\right) \leq n\exp\left(\frac{-m}{8}\min\left\{(\mu_{\max} + t - 1)^2, \mu_{\max} + t - 1\right\}\right).$$

To ensure that the random variable on the left hand side of (2.22) exceeds the one on the right hand side with large probability, we can substitute the bounds we derived for each term, and check when the inequality holds. This results (up to some constant loss in the δ factor) in a condition on the problem parameters under which (2.22) holds for

a fixed $(u, u') \in \mathcal{S} \times \mathcal{S}^c$. Applying a union bound over all $k(d-k)$ pairs gives the final requirement on n.

Note that the leading terms on the right hand side of (2.22) would roughly be $\sqrt{(k^2/m^2 n)\log d/\delta}$ or $\sqrt{(k/mn)\log d/\delta}$ (assuming $m \geq 2\log(d/\delta)$, see the proof in Section 2.6.2 for details), while the left hand side would roughly be a constant, leading to the following result.

Lemma 2.4.6. *The separation condition* (2.22) *holds for every* $(u, u') \in \mathcal{S} \times \mathcal{S}^c$*, with probability at least* $1-\delta$*, provided* $m \geq 2\log(d/\delta)$ *and*

$$n \geq c\frac{x_{max}^4}{x_{\min}^4} \max\left\{\left(\frac{k}{m} + \frac{\sigma^2}{x_{\max}^2}\right)\log\frac{d}{\delta}, \left(\frac{k}{m} + \frac{\sigma^2}{x_{\max}^2}\right)^2 \log\frac{d}{\delta}\right\},$$

for an absolute constant c*.*

By defining $\mathcal{E}$ as the event that the measurement matrices satisfy condition (2.22) for every $(u, u') \in \mathcal{S} \times \mathcal{S}^c$, we can see that the probability of error in (2.20) is at most δ, provided n satisfies the condition in Lemma 2.3.8. This completes the proof of Theorem 2.4.1.

We make a few observations before moving to simulation results. The squaring step in (2.19) in the variance-based estimator is done to ensure that the averaging does not lead to cancellations for coordinates that lie in the true support. In fact, if the inputs are all nonnegative, then a mean-based estimator would suffice, and it would lead to a smaller sample complexity upper bound. In particular, inputs with both positive and negative values lead to the increased sample complexity of $k^2/m^2 \cdot \log d$, as we will see in the lower bound result also where Gaussian inputs constitute the difficult case.

2.5 Simulation results

In this section, we numerically evaluate the performance of the closed-form estimator in (2.3). Our focus will be on exact support recovery and we will study the performance of our estimator over multiple trials. For our experiments, we use measurement matrices

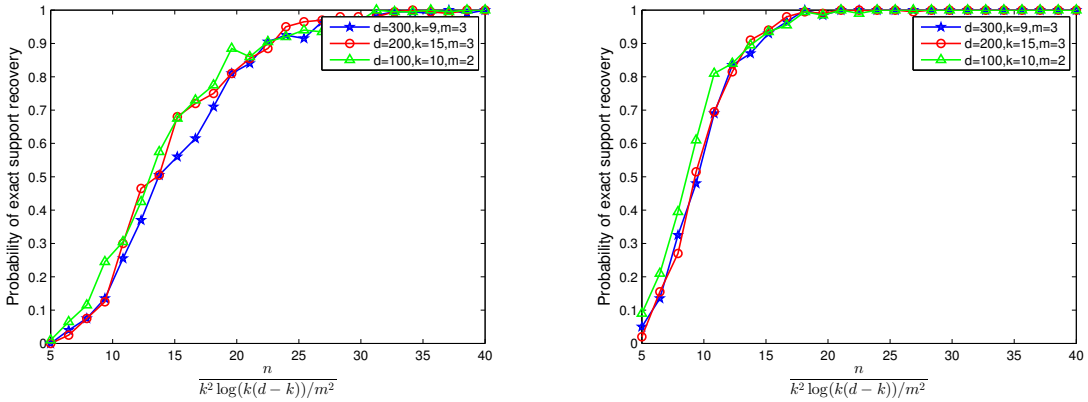


Figure 2.2: Support recovery performance of the closed-form estimator for Gaussian and Rademacher priors.

that are independent across samples and have i.i.d. $\mathcal{N}(0, 1/m)$ entries. To generate measurements, we first pick a support uniformly at random from all possible supports of size k. Next, the data vectors are generated according to one of two methods. In the first method, the nonzero entries of the data have i.i.d. $\mathcal{N}(0, 1)$ entries. In the second method, the nonzero entries are i.i.d Rademacher (i.e., $\{+1, -1\}$-valued with equal probability). Both these distributions are subgaussian with variance parameters that are a constant multiple of the respective variances. We generate noiseless measurements Y_1^n according to the linear model described before. For a fixed value of d, k, m and n, we generate multiple instances of the problem and provide it as input to the estimator. For every instance, we declare success or failure depending on whether the support is exactly recovered or not and the success rate is the fraction of instances on which the recovery is successful. For our experiments, we performed 200 trials for every set of parameters. We can see from Figure 2.2 that the experimental results closely agree with our predictions. Also, the constant of proportionality is small, roughly between 15 and 20. We also perform simulations for the case when the measurements are noisy. In particular, we consider noise vectors $W_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$ for different values of σ^2, and X_1^n Gaussian distributed as described before. We plot the probability of exact support recovery against the normalized number of samples for different noise levels, while the other parameters are kept fixed at $d = 100$, $m = 2$, and $k = 10$. It can be seen from Figure 2.3 that the four

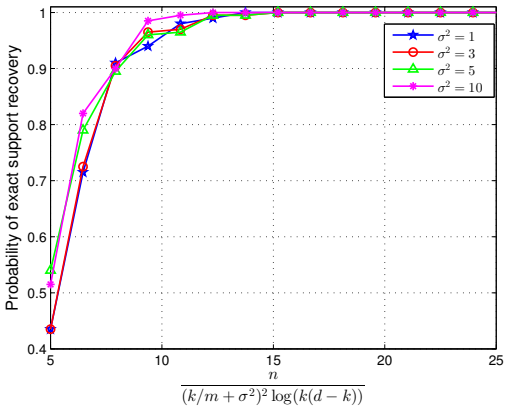


Figure 2.3: Performance of the closed-form estimator for different noise levels with $d = 100,\ m = 2,\ k = 10$.

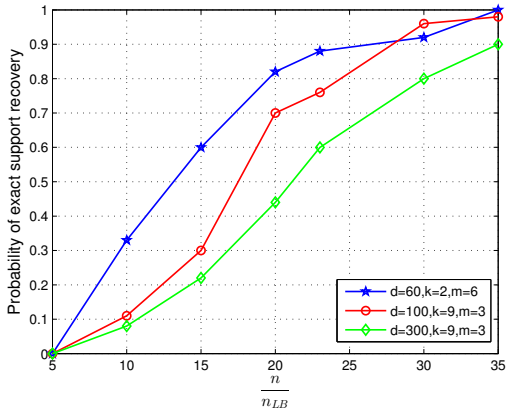


Figure 2.4: Performance of MSBL in the noiseless case for different parameter values.

curves overlap, indicating that the scaling of n with respect to the noise variance is tight. Finally, Figure 2.4 shows the performance of MSBL [88], where we plot the probability of exact support recovery against the normalized number of samples (the normalization factor $n_{LB} = (k^2(1 - m/k)^4/m^2)\log(k(d-k))$ is from the lower bound established in the next chapter). It can be seen that the curves do not overlap, indicating that MSBL has a different scaling of n with respect to the parameters m, k, d than what is obtained by our lower bound.

2.6 Remaining proofs

2.6.1 Proofs from Section 2.3.1

Proof of Lemma 2.3.5. By independence of $\{X_i\}_{i=1}^n$ and $\{\Phi_i\}_{i=1}^n$, we have that

$$\mathtt{var}_{X,\Phi}\left(\tilde{\lambda}_i\right) = \frac{1}{n^2}\sum_{j=1}^{n} \mathtt{var}_{X,\Phi}\left(X_j^\top B_{ji} X_j\right). \tag{2.25}$$

where $B_{ji} = \Phi_j^\top \Phi_{ji} \Phi_{ji}^\top \Phi_j$. Consider one term in the summation and let $X_j^\top B_{ji} X_j \equiv X^\top B_i X$. The variance calculation is based on the following useful decomposition

$$\texttt{var}_{X,\Phi}\left(X^\top B_i X\right) = \mathbb{E}_\Phi\left[\texttt{var}_X\left(X^\top B_i X|\Phi\right)\right] + \texttt{var}_\Phi\left(\mathbb{E}_X\left[X^\top B_i X|\Phi\right]\right). \tag{2.26}$$

We first look at the term $\texttt{var}_\Phi\left(\mathbb{E}_X\left[X^\top B_i X|\Phi\right]\right)$ and note that

$$\begin{aligned}\mathbb{E}_X\left[X^\top B_i X|\Phi\right] &= \texttt{Tr}\left(\mathbb{E}_X\left[B_i X X^\top|\Phi\right]\right)\\ &= \texttt{Tr}\left(B_i K_\lambda\right)\\ &= \sum_{j\in\mathcal{S}}(B_i)_{jj}.\end{aligned} \tag{2.27}$$

Now, note that

$$\sum_{j\in\mathcal{S}}(B_i)_{jj} = \begin{cases}\|\Phi_i\|_2^4 + \sum_{j\in\mathcal{S}\setminus\{i\}}(\Phi_j^\top\Phi_i)^2, \text{ if } i\in\mathcal{S},\\ \sum_{j\in\mathcal{S}}(\Phi_j^\top\Phi_i)^2, \text{ otherwise,}\end{cases} \tag{2.28}$$

This gives

$$\begin{aligned}\texttt{var}_\Phi\left(\mathbb{E}\left[X^\top B_i X|\Phi\right]\right) &= \texttt{var}\left(\sum_{j\in\mathcal{S}}(B_i)_{jj}\right)\\ &= \begin{cases}\texttt{var}\left(\|\Phi_i\|_2^4 + \sum_{j\in\mathcal{S}\setminus\{i\}}(\Phi_j^\top\Phi_i)^2\right), \text{ if } i\in\mathcal{S},\\ \texttt{var}\left(\sum_{j\in\mathcal{S}}(\Phi_j^\top\Phi_i)^2\right), \text{ otherwise.}\end{cases}\end{aligned} \tag{2.29}$$

Subsequent calculations mostly rely on moments of inner products and norms of Gaussian random vectors which are stated in Section 2.6.3. In particular, consider $\texttt{var}\left(\sum_{j\in\mathcal{S}}(\Phi_j^\top\Phi_i)^2\right)$ for the $i\in\mathcal{S}^c$ case in (2.29). We have using Lemma 2.6.12,

$$\begin{aligned}\texttt{var}\left(\sum_{j\in\mathcal{S}}(\Phi_j^\top\Phi_i)^2\right) &= \mathbb{E}\left[\left(\sum_{j\in\mathcal{S}}(\Phi_j^\top\Phi_i)^2\right)^2\right] - \left(\mathbb{E}\left[\sum_{j\in\mathcal{S}}(\Phi_j^\top\Phi_i)^2\right]\right)^2\\ &= \frac{k^2}{m^2} + \frac{2k^2}{m^3} + \frac{2k}{m^2} + \frac{4k}{m^3} - \frac{k^2}{m^2}\end{aligned}$$

$$= \frac{2k}{m^2} + \frac{4k}{m^3} + \frac{2k^2}{m^3}. \tag{2.30}$$

We now consider the $i \in \mathcal{S}$ case. Using Lemma 2.6.14, we get

$$\texttt{var}\left(\|\Phi_i\|_2^4 + \sum_{j \in \mathcal{S}\setminus\{i\}} (\Phi_j^\top \Phi_i)^2 \right) = \frac{8}{m} + \frac{12}{m^2} + \frac{50}{m^3} + \frac{10k}{m^2} + \frac{2k^2}{m^3} + \frac{16k}{m^3}. \tag{2.31}$$

Finally, from (2.29), (2.30) and (2.31), we have

$$\texttt{var}_\Phi \left(\mathbb{E}\left[X^\top B_i X | \Phi \right] \right) = \begin{cases} \frac{8}{m} + \frac{12}{m^2} + \frac{50}{m^3} + \frac{10k}{m^2} + \frac{2k^2}{m^3} + \frac{16k}{m^3}, \text{ if } i \in \mathcal{S}, \\ \frac{2k}{m^2} + \frac{2k^2}{m^3} + \frac{4k}{m^3}, \text{ otherwise.} \end{cases} \tag{2.32}$$

We now compute the first term in (2.26), $\mathbb{E}_\Phi \left[\texttt{var}(X^\top B_i X | \Phi) \right]$. Note that

$$\begin{aligned} \texttt{var}(X^\top B_i X | \Phi) &= \texttt{var}(X_{\mathcal{S}}^\top (B_i)_{\mathcal{S},\mathcal{S}} X_{\mathcal{S}} | \Phi) \\ &= 2 \ \texttt{Tr}((B_i)_{\mathcal{S},\mathcal{S}}^2), \end{aligned} \tag{2.33}$$

where the second step follows from Lemma 2.6.15. Let us consider the $i \in \mathcal{S}$ case first. Also, for ease of notation, let $\mathcal{S} = \{1, \ldots, k-1\} \cup \{i\}$. Then,

$$\begin{aligned} \mathbb{E}_\Phi \left[\texttt{var}(X^\top B_i X | \Phi) \right] =& 2\mathbb{E} \left[\|\Phi_i\|^8 + \sum_{j \in \mathcal{S}\setminus\{i\}} (\Phi_j^\top \Phi_i)^4 + 2 \sum_{u \in \mathcal{S}\setminus\{i\}} \|\Phi_u\|^4 (\Phi_u^\top \Phi_i)^2 \right] \\ &+ \mathbb{E} \left[2 \sum_{\substack{v,w \in \mathcal{S}\setminus\{i\} \\ v \neq w}} (\Phi_v^\top \Phi_i)^2 (\Phi_w^\top \Phi_i)^2 \right]. \end{aligned} \tag{2.34}$$

Using Lemmas 2.6.13 and 2.6.12, we get for $i \in \mathcal{S}$,

$$\begin{aligned} \mathbb{E}\left[\texttt{var}(X^\top B_i X | \Phi) \right] &= 2\mathbb{E}\left[\texttt{Tr}((B_i)_{\mathcal{S},\mathcal{S}}^2) \right] \\ &= 2\Bigg[\left(1 + \frac{12}{m} + \frac{44}{m^2} + \frac{48}{m^3} \right) + (k-1) \left(\frac{3}{m^2} + \frac{6}{m^3} \right) \end{aligned}$$

$$+ 2(k-1)\left(\frac{1}{m} + \frac{6}{m^2} + \frac{8}{m^3}\right) + 2(k-1)(k-2)\left(\frac{1}{m^2} + \frac{2}{m^3}\right)\Bigg] \tag{2.35}$$

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

$$\begin{aligned} 2\mathtt{Tr}((B_i)^2_{\mathcal{S},\mathcal{S}}) &= 2\sum_{j\in\mathcal{S}}(\Phi_j^\top \Phi_i)^4 + 4\sum_{\substack{v,w\in\mathcal{S}\\ v\neq w}}(\Phi_v^\top\Phi_i)^2(\Phi_w^\top\Phi_i)^2 \\ &= 2k\left(\frac{3}{m^2} + \frac{6}{m^3}\right) + 4k(k-1)\left(\frac{2}{m^3} + \frac{1}{m^2}\right), \end{aligned}$$

which gives

$$\mathbb{E}\left[\mathtt{var}(X^\top B_i X|\Phi)\right] = \frac{2k}{m^2} + \frac{4k}{m^3} + \frac{8k^2}{m^3} + \frac{4k^2}{m^2}. \tag{2.36}$$

Thus, from (2.32), (2.35) and (2.36), we have the variance in (2.26):

$$\mathtt{var}(\tilde{\lambda}_i) \leq \begin{cases} \frac{c}{n}\left(1 + \frac{k}{m} + \frac{k^2}{m^2} + \frac{k}{m^2} + \frac{k^2}{m^3}\right), \text{ if } i \in \mathcal{S} \\ \frac{c'}{n}\left(\frac{k}{m^2} + \frac{k^2}{m^3} + \frac{k^2}{m^2}\right), \text{ otherwise.} \end{cases}, \tag{2.37}$$

where c and c' are absolute constants. □

We recall the statement of Lemma 2.3.8 here for easy reference.

Lemma 2.6.1. *For all pairs $(i, i') \in S \times S^c$, the separation condition*

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^n \alpha_{ji}^2 - \frac{1}{n}\sum_{j=1}^n \alpha_{ji'}^2 \geq &\max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^n \alpha_{ji}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2\log\frac{1}{\delta'}\right\} \\ &+ \max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^n \alpha_{ji'}^4 \log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji'}^2\log\frac{1}{\delta'}\right\} \end{aligned} \tag{2.38}$$

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

Proof. The proof involves studying the tail behaviour of each term in (2.38). In particular, we derive a lower bound on the first term and upper bounds on the remaining terms that

hold with high probability over the subgaussian measurement ensemble, and establish conditions under which the separation in (2.38) holds for a fixed pair (i, i'). A union bound over all $k(d-k)$ pairs then gives us the result claimed in the lemma. The key technicality is to keep track of the leading term that is contributed by each of the terms in (2.38). To get a rough idea of the behaviour of these terms, recall that α_{ji} depends on inner products between the columns Φ_j. In particular, it involves the sum of $O(k)$ inner product squared terms, and this scales as $O(k/m)$ in expectation. The right side in (2.38) thus effectively leads to a $O(1)$ term. The left side, on the other hand, gives rise to a $O(\sqrt{1/n \cdot (k/m)^2})$ term in expectation, leading to the overall requirement of $n = O(k^2/m^2)$. In what follows, we make these arguments precise using tail bounds for each of the terms in (2.38).

For clarity of presentation, details of the tail bounds for each term in (2.38) are presented in Section 2.6.3, which in turn build on standard concentration bounds for subgaussian and subexponential random variables reviewed as preliminaries in Section 2.6. Also, while analyzing each term in (2.38), we use the same symbol μ to denote the expectation of that term to keep notation simple. Similarly, the definitions of terms like μ_1, μ_2 and μ_3 will be clear from the context.

For the first term on the left side of (2.38), we study the behaviour of its left tail. That is, we look at $\Pr\left(\frac{1}{n}\sum_{j=1}^{n} \alpha_{ji}^2 \leq \mu - t\right)$, where recall

$$\alpha_{ji}^2 = \|\Phi_{ji}\|_2^4 + \sum_{l \in S \setminus \{i\}} (\Phi_{jl}^\top \Phi_{ji})^2 + \sigma^2 \|\Phi_{ji}\|_2^2, \tag{2.39}$$

and $\mu = \mathbb{E}\left[(1/n)\sum_{j=1}^{n} \alpha_{ji}^2\right]$. Further, let μ_1, μ_2 and μ_3 denote the mean of each of the three terms. By a union bound argument, it suffices to bound the normalized sum of each of the three terms separately. Notice that all these terms essentially depend on the lengths of the columns or the inner products between the columns of the measurement matrix, and our goal will be obtain concentration bounds for these terms. While $\|\Phi_{ji}\|_2^2$ is clearly subexponential, $\|\Phi_{ji}\|_2^4$ and $(\Phi_{jl}^\top \Phi_{ji})^2$ have heavier tails. In Section 2.6.3, we provide results on the tail behaviour of these terms. Using Lemma 2.6.7, we have for any

$t > 0$,

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

for

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

Further, from Lemma 2.6.9, we have that when $n \ge (c_2^2/c_1)\log(12/\varepsilon)$,

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

for

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

Finally, from Lemmas 2.6.4 and 2.6.5, we can see that $\|\Phi_{ji}\|_2^2$ is subexponential with parameters $(c_1/m, c_2/m)$ and that $(\sigma^2/n)\sum_{j=1}^{n}\|\Phi_{ji}\|_2^2$ is subexponential with parameters $(c_1\sigma^4/mn, c_2\sigma^2/mn)$. Using the subexponential concentration bound from Lemma 2.6.6 gives

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

for

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

Combining these results using a union bound step, we see that

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

for

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

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

We now consider the second term on the left side of (2.38), and observe that it consists of terms similar to the ones we encountered in the previous calculation. Our focus will be on the *right* tail this time, i.e., we will study $\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\alpha_{ji'}^2 \geq \mu + t\right)$ for $i' \in S^c$, where

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

We use Lemma 2.6.6 to get

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

for

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

and Lemma 2.6.10 to get

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

for

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

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

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

for

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

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

For the third term in (2.38), namely, $\max\left\{\sqrt{\frac{c_1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4\log\frac{1}{\delta'}}, \frac{c_2}{n}\max_{j\in[n]}\alpha_{ji}^2\log\frac{1}{\delta'}\right\}$, we consider the possibility of either argument attaining the maximum and study the respective right tails.

First, we look at $\Pr\left(\sqrt{\frac{1}{n^2}\sum_{j=1}^{n}\alpha_{ji}^4} \geq \mu + t\right)$ for $i \in S$. We note that by the union bound,

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

$$+ n\Pr\left(\sigma^2\|\Phi_{1i}\|_2^2 \geq \frac{\sqrt{n}}{3}(\mu+t)\right). \tag{2.40}$$

We use Lemma 2.6.3 for the first and third terms and Lemma 2.6.10 for the second term. A direct application of Lemma 2.6.10 with $n = 1$ for the second term however requires the assumption that $m \geq (c_2^2/c_1)\log(12n/\varepsilon)$ (note that the second term in (2.40) needs to be upper bounded by $\varepsilon/3n$). While in our setting such an assumption on n is acceptable, we would like to avoid making this assumption on m at this stage. We therefore omit the simplification done at the end of Lemma 2.6.10 to get

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

for

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

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

$$\begin{aligned}
\Pr\left(\max_{j\in[n]}\alpha_{ji}^2 \geq \mu + t\right) &\leq \sum_{j=1}^{n}\Pr\left(\alpha_{ji}^2 \geq \mu + t\right) \\
&\leq \sum_{j=1}^{n}\left[\Pr\left(\|\Phi_{ji}\|_2^4 \geq \frac{\mu+t}{3}\right) + \Pr\left(\sum_{l\in S\setminus\{i\}}(\Phi_{ji}^\top\Phi_{jl})^2 \geq \frac{\mu+t}{3}\right)\right. \\
&\quad \left. + \Pr\left(\sigma^2\|\Phi_{ji}\|_2^2 \geq \frac{\mu+t}{3}\right)\right].
\end{aligned} \tag{2.41}$$

We now handle each of the three terms on the right-side of (2.6.1) separately. We will use Lemma 2.6.6 for the first and third terms and Lemma 2.6.9 for the second term. In particular, for every $j \in [n]$, we have that

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

for

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

and that

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

for

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

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

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

for

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

Substituting those bounds into (2.6.1), we get

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

for

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

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

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

for

$$\mu + t = \frac{3}{n}(1 + f(m, \varepsilon/3n)) \cdot \max\left\{\sigma^2, 1 + f(m, \varepsilon/3n), \frac{k-1}{m} + \sqrt{c_2 \frac{k-1}{m^2} \log \frac{9n}{\varepsilon}}\right\}. \quad (2.42)$$

Comparing (2.41) and (2.42), we see that $\frac{1}{n} \max_{j \in [n]} \alpha_{ji}^2$ is $O(k/mn + \sigma^2/n)$ which decays faster with respect to n compared to $\sqrt{\frac{1}{n^2} \sum_{j=1}^{n} \alpha_{ji}^4}$, which is $O(k/m\sqrt{n} + \sigma^2/\sqrt{n})$. Thus, the third term in (2.38) is dominated by the $O(k/m\sqrt{n} + \sigma^2/\sqrt{n})$ term, which is what we retain in our subsequent calculations.

Finally, for the fourth term in (2.38), we first look at $\Pr\left(\sqrt{(1/n^2) \sum_{j=1}^{n} \alpha_{ji'}^4} \geq \mu + t\right)$ for $i' \in S^c$. Using similar arguments as in the previous calculation, we get

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

for

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

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

The foregoing calculations provide bounds on each of the four terms occuring in (2.38),

that hold with high probability. We note that the left-side of (2.38) is lower bounded by

$$1 - \frac{1}{m} - 2\sqrt{\frac{c_1}{mn}\log\frac{24}{\varepsilon}}\left(1 + \sigma^2 + \frac{k}{m}\right) + \frac{c_1}{mn}\log\frac{6}{\varepsilon} \tag{2.43}$$

$$-2\sqrt{\frac{c_1 k}{m^2 n}\log\frac{24}{\varepsilon}}\left(1 + \frac{c_2}{m}\log\frac{24n}{\varepsilon}\right) \tag{2.44}$$

with probability at least $1-\varepsilon$, and that the right-side is upper bounded by

$$\begin{aligned}
&5\sqrt{\frac{c_1}{n}\log\frac{1}{\delta'}}\Bigg[\left(1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{6n}{\varepsilon}}, \frac{c_2}{m}\log\frac{6n}{\varepsilon}\right\}\right)^2 \\
&+ \left(\frac{k}{m} + \max\left\{\frac{c_2}{m}\log\frac{18n}{\varepsilon}, \sqrt{c_1\frac{k}{m^2}\log\frac{18n}{\varepsilon}}\right\}\right)\left(1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{18n}{\varepsilon}}, \frac{c_2}{m}\log\frac{18n}{\varepsilon}\right\}\right) \\
&+ \sigma^2\left(1 + \max\left\{\sqrt{\frac{c_1}{m}\log\frac{6n}{\varepsilon}}, \frac{c_2}{m}\log\frac{6n}{\varepsilon}\right\}\right)\Bigg],
\end{aligned} \tag{2.45}$$

with probability at least $1-\varepsilon$. To ensure that (2.38) holds with probability at least $1-\varepsilon$ for a fixed $(i, i') \in S \times S^c$, we need that (2.43) exceeds (2.6.1). For further simplification, we assume m to be sufficiently large to handle the $\log n$ terms. This assumption on m can possibly be removed by handling the sum in Lemma 2.6.8 and (2.40) directly and not using the union bound. Choosing $\varepsilon = \delta/(4\ k(d-k))$ to account for the union bound over all (i, i') pairs and focusing on the $n = O((k/m + 1 + \sigma^2)^2 \log(1/\delta'))$ regime, we see that (2.43) exceeds (2.6.1) and separation holds if[3] $m \geq (\log k)^2$.

Thus,

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

samples suffice to ensure separation between the typical values and to guarantee that (2.38) holds with probability at least $1-\delta$. □

[3]We use this condition to show that $(1/\sqrt{m})\log(k/m) \leq 1$ and the dominating term on the right-side of (2.6.1) is k/m.

2.6.2 Proofs from Section 2.4

Proof of Lemma 2.4.3. We first note that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (\|\Phi_{iu}\|_2^6 - \mathbb{E}\left[\|\Phi_{iu}\|_2^6\right]) \right\|_{\mathcal{L}_p} = \frac{1}{nm^3} \left\| \sum_{i=1}^{n} (V_i^3 - \mathbb{E}\left[V_i^3\right]) \right\|_{\mathcal{L}_p}, \tag{2.46}$$

where $V_i \stackrel{\text{def}}{=} m\|\Phi_{iu}\|_2^2 \sim \chi_m^2$, and χ_m^2 denotes the chi-square distribution with m degrees of freedom. To bound the moment of the sum, we use the following form of Rosenthal's inequality stated in [56].

Lemma 2.6.2 ([56])**.** *Let $Z_1, \ldots, Z_n$ be independent and identically distributed random variables with mean zero. Then, for every $p \geq 2$,*

$$\left\| \sum_{i=1}^{n} Z_i \right\|_{\mathcal{L}_p} \leq c\left(pn^{\frac{1}{p}} \|Z_1\|_{\mathcal{L}_p} + \sqrt{pn} \|Z_1\|_{\mathcal{L}_2} \right),$$

for an absolute constant c.

In view of Lemma 2.6.2, we now upper bound the $\mathcal{L}_p$ norm of each summand on the right side of (2.46) as follows:

$$\begin{aligned}
\|V_i^3 - \mathbb{E}\left[V_i^3\right]\|_{\mathcal{L}_p} &\leq \|V_i^3\|_{\mathcal{L}_p} + \mathbb{E}\left[V_i^3\right] \\
&= (\mathbb{E}\left[V_i^{3p}\right])^{\frac{1}{p}} + \mathbb{E}\left[V_i^3\right] \\
&= \left(2^{3p} \frac{\Gamma(3p + m/2)}{\Gamma(m/2)} \right)^{\frac{1}{p}} + 2^3 \frac{\Gamma(3 + m/2)}{\Gamma(m/2)} \\
&\leq 2^3 \left(e^{\frac{1}{p}} (3p + m/2)^3 + e(3 + m/2)^3 \right) \\
&\leq 2^6 (3p + m/2)^3,
\end{aligned}$$

where we used the fact that $V_i \sim \chi_m^2$ in the third step and $\Gamma(x + a)/\Gamma(x) \leq e(x + a)^a$ for all $x \geq 1$, $a > 0$ in the fourth step. Together with Lemma 2.6.2, this yields for $p \geq 2$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (\|\Phi_{iu}\|_2^6 - \mathbb{E}\left[\|\Phi_{iu}\|_2^6\right]) \right\|_{\mathcal{L}_p} \leq \frac{c2^6}{nm^3} \left(pn^{\frac{1}{p}} (3p + m/2)^3 + \sqrt{pn} (6 + m/2)^3 \right)$$

$$\leq c2^6\left(\frac{p}{n^{1-\frac{1}{p}}}\max\left\{1,\frac{(6p)^3}{m^3}\right\}+7^3\sqrt{\frac{p}{n}}\right)$$
$$\leq c'\max\left\{\frac{p}{n^{1-\frac{1}{p}}},\frac{p^4}{m^3n^{1-\frac{1}{p}}},\sqrt{\frac{p}{n}}\right\}. \tag{2.47}$$

Note that from (2.23), we expect p to be of the form $n^{c''}$ for some constant c'', in which case $p/n^{1-\frac{1}{p}} = (p/n)e^{\frac{1}{ec''}}$. We focus on this regime, and obtain using (2.24) and (2.47),

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(\|\Phi_{iu}\|_2^6-\mathbb{E}\left[\|\Phi_{iu}\|_2^6\right])\geq t\right|\right)\leq\exp\left(-C\min\left\{nt,(m^3nt)^{\frac{1}{4}},nt^2\right\}\right),$$

for every $t>0$. □

Proof of Lemma 2.4.5. Let $\mu_{\max}=\max_{i\in[n]}\|\Phi_{iu}\|_2^2$. The proof follows by noting that for every $t>0$,

$$\Pr\left(\max_{i\in[n]}\|\Phi_{iu}\|_2^2\geq\mu_{\max}+t\right)\leq\sum_{i=1}^{n}\Pr\left(\|\Phi_{iu}\|_2^2-1\geq t'\right),$$

where $t'=\mu_{\max}+t-1$, and using the fact that $m\|\Phi_{iu}\|_2^2\sim\chi_m^2$ to get

$$\Pr\left(\max_{i\in[n]}\|\Phi_{iu}\|_2^2\geq t\right)\leq\exp\left(-\frac{m}{8}\min\{t'^2,t'\}\right).$$

□

Proof of Lemma 2.4.6. The proof involves finding upper and lower bounds, respectively, on the left-hand side and right-hand side of (2.22) that hold with high probability, and then simplifying to obtain the condition on n stated in the lemma. Note that there are two probability of error parameters here, one from the criterion in (2.20), and another required for (2.22). To avoid confusion, will use δ for the former and δ' for the latter (we will eventually set $\delta'=\delta/(k(d-k))$).

For the left-hand side of (2.22), it follows from Lemma 2.4.4 that

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\|\Phi_{iu}\|_2^4\geq 1+\frac{2}{m}-t\right)\leq\delta',$$

when

$$t \geq \frac{1}{C} \max \left\{ \frac{1}{n} \log \frac{1}{\delta'}, \frac{1}{nm^2} \left(\log \frac{1}{\delta'} \right)^3, \sqrt{\frac{1}{n} \log \frac{1}{\delta'}} \right\},$$

where the maximum in th expression above is the third term provided $m > \log(1/\delta')$. Further, since $m\|\Phi_{iu}\|_2^2 \sim \chi_m^2$, we have

$$\Pr \left(\frac{1}{mn} \sum_{i=1}^{n} \|\Phi_{iu}\|_2^2 \geq \frac{1}{m} - t \right) \leq \delta',$$

when

$$t \geq \frac{2}{m} \left(\sqrt{\frac{1}{mn} \log \frac{1}{\delta'}} + \frac{1}{mn} \log \frac{1}{\delta'} \right).$$

It follows that the left-hand side of (2.22) is at least $cx_{\min}^2/x_{\max}^2$ with probability at least $2\delta'$, for an absolute constant c, provided $m > \log(1/\delta')$.

We now proceed to find a high probability upper bound on the right-hand side of (2.22). Lemmas 2.4.3 and 2.4.4 can be used to upper bound the first three terms, and Lemma 2.4.5 can be used for the last term. In particular, we have

$$\Pr \left(\frac{1}{n} \sum_{i=1}^{n} \|\Phi_{iu}\|_2^6 \geq 1 + \frac{6}{m} + \frac{8}{m^2} + t \right) \leq \delta',$$

when

$$t \geq \frac{1}{C} \max \left\{ \frac{1}{n} \log \frac{1}{\delta'}, \frac{1}{nm^3} \left(\log \frac{1}{\delta'} \right)^4, \sqrt{\frac{1}{n} \log \frac{1}{\delta'}} \right\}.$$

Further,

$$\Pr \left(\frac{1}{n} \sum_{i=1}^{n} \|\Phi_{iu}\|_2^4 \geq 1 + \frac{2}{m} + t \right) \leq \delta',$$

when

$$t \geq \frac{1}{C} \max\left\{\frac{1}{n}\log\frac{1}{\delta'}, \frac{1}{nm^2}\left(\log\frac{1}{\delta'}\right)^3, \sqrt{\frac{1}{n}\log\frac{1}{\delta'}}\right\},$$

and

$$\Pr\left(\max_{i\in[n]} \|\Phi_{iu}\|_2^2 \geq 1 + \max\left\{\sqrt{\frac{8}{m}\log\frac{n}{\delta'}}, \frac{8}{m}\log\frac{n}{\delta'}\right\}\right) \leq \delta'.$$

To simplify the right-hand side of (2.22), note that after substituting the bounds above, the leading terms arise from the mean of Φ_i dependent terms (i.e. normalized sum and the normalized maximum), which is roughly 1. In particular, we see that the leading terms are roughly $k/mn \cdot \log(1/\delta)$ and $k^2/m^2 \cdot \log(d/\delta)$, provided $m \geq \log(1/\delta')$ (this condition ensures that the deviation terms for the Φ_i dependent terms are small). Using this observation and recalling that the left-hand side in (2.22) is a constant gives, after simplification, that (2.22) holds for a fixed $(u, u') \in \mathcal{S} \times \mathcal{S}^c$ with probability at least $1-\delta'$, provided $m \geq \log(1/\delta')$ and

$$n \geq c\frac{x_{max}^4}{x_{\min}^4} \max\left\{\left(\frac{k}{m} + \frac{\sigma^2}{x_{\max}^2}\right)\log\frac{d}{\delta}, \left(\frac{k}{m} + \frac{\sigma^2}{x_{\max}^2}\right)^2 \log\frac{d}{\delta}\right\},$$

for an absolute constant c. We now apply a union bound over all pairs (u, u') and choose $\delta' = \delta/(k(d-k))$. Finally, noting that $\log(1/\delta') \leq 2\log(d/\delta)$, gives us the result stated in the lemma. □

2.6.3 Useful lemmas

Definition 2.6.1. *A random variable X is subgaussian with variance parameter σ^2, denoted $X \sim subG(\sigma^2)$, if*

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

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

Definition 2.6.2. *A random variable X is subexponential with parameters σ^2 and $b > 0$,*

denoted $X \sim subexp(\sigma^2, b)$*, if*

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

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

Lemma 2.6.3. *Let* X *be a subexponential random variable with parameters* v^2 *and* $b > 0$ *(denoted* $X \sim subexp(v^2, b)$*), that is,*

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

Then,

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

Proof. See [82, Proposition 2.2]. □

Lemma 2.6.4. *Let* $X \sim subG(\sigma^2)$ *with* $\mathbb{E}\left[X\right] = 0$*. Then* $X^2 \sim subexp(128\sigma^4, 8\sigma^2)$*.*

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

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

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

Now, for $X \sim \text{subG}(\sigma^2)$, we have the following upper bound on the moments of X from [13, Theorem 2.1] : $\mathbb{E}\left[X^{2q}\right] \leq 2q!2^q\sigma^{2q}$. This gives

$$\mathbb{E}\left[e^{\theta(Y-\mathbb{E}[Y])}\right] \leq 1 + 2\sum_{q=2}^{\infty} \frac{\theta^q}{q!} q!2^{2q}\sigma^{2q}$$

$$= 1 + \frac{32\theta^2\sigma^4}{1 - 4\theta\sigma^2}, \quad \theta < \frac{1}{4\sigma^2}.$$

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

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

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

Lemma 2.6.5. *Let $X_i \sim subexp(v_i^2, b_i)$ be independent subexponential random variables for $i \in [n]$. Then, for a constant $a \in \mathbb{R}$, we have that $aX_1 \sim subexp(a^2v_1^2, |a|b_1)$ and $\sum_{i=1}^n X_i \sim subexp(\sum_{i=1}^n v_i^2, max_{i\in[n]}b_i)$.*

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

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

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

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

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

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

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

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

Lemma 2.6.6. *Let $Z_1, \ldots, Z_n$ be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Then, there exist absolute*

constants c_1 and c_2 such that for any $t > 0$,

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

Equivalently, for any $\epsilon > 0$,

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

Proof. Since $Z_{jl} \sim \text{subG}(1/m)$ for any $j \in [n]$ and $l \in [m]$, we have from Lemma 2.6.4 that $Z_{jl}^2 \sim \text{subexp}(c_1/m^2, c_2/m)$ for some absolute constants c_1 and c_2. Using properties of subexponential random variables from Lemma 2.6.5, we can show that the normalized sum $\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2$ is also subexponential with parameters $(c_1/mn, c_2/mn)$. Noting that $\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2\right] = 1$ and using the tail bound from Lemma 2.6.3 we get for $t > 0$,

$$\Pr\left(\left|\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2 - 1\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{mn}{c_1}t^2, \frac{mn}{c_2}t\right\}\right). \tag{2.48}$$

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

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

Substituting the above into (2.48) gives us the result. □

Lemma 2.6.7. *Let $Z_1, \ldots, Z_n$ be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Then, there exist absolute constants c_1 and c_2 such that for any $\varepsilon > 0$,*

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

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

we have

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

where $t' = 1 - \sqrt{\mu - t}$. Using Lemma 2.6.6 and reparameterizing with respect to $\varepsilon > 0$ gives the result.

□

Lemma 2.6.8. *Let $Z_1, \ldots, Z_n$ be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Then, there exist absolute constants c_1 and c_2 such that for any $\varepsilon > 0$,*

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

Proof. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^4\right]$ and note as we did in Lemma 2.6.6 that $\|Z_j\|_2^2 \sim \text{subexp}(c_1/m, c_2/m)$. We have by union bound that

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

where the last inequality follows from Lemma 2.6.3 with $t' = \sqrt{\mu + t} - 1$. Equating the expression on the right to ε and reparameterizing gives the result.

□

Lemma 2.6.9. *Let $Z_j, Y_{j1}, \ldots, Y_{j,k-1}$, $j \in [n]$, be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2\right]$. Then, there exist absolute constants c_1 and c_2 such that*

for any $\varepsilon > 0$,

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

for

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

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

Proof. Note that conditioned on Z_j, the random variable $Y_{jl}^{\top}Z_j$ is subgaussian with parameter $\|Z_j\|_2^2/m$, for any $j \in [n]$ and $l \in [k-1]$. Using Lemmas 2.6.4 and 2.6.5, we have that the normalized sum $(1/n)\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^{\top}Z_j)^2$, conditioned on $\{Z_j\}_{j=1}^n$, is subexponential with parameters v^2 and b where

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

Let $\mu' = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^{\top}Z_j)^2\middle|\{Z_j\}_{j=1}^n\right]$. Since the variance and the variance parameter are equal, we have

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

From Lemma 2.6.3 we have that for $t > 0$,

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^{\top}Z_j)^2 \le \mu - t\middle|\{Z_j\}_{j=1}^n\right)\\
&= \Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^{\top}Z_j)^2 - \mu' \le \mu - t - \mu'\middle|\{Z_j\}_{j=1}^n\right)\\
&\le \exp\left(-\min\left\{\frac{m^2n^2(t')^2}{c_1(k-1)\sum_{j=1}^{n}\|Z_j\|_2^4}, \frac{mnt'}{c_2\max_{j\in[n]}\|Z_j\|_2^2}\right\}\right) \qquad (2.49)
\end{aligned}$$

where $t' = \mu' + t - \mu$. We now handle the Z_j-dependent terms in the exponent. In

particular, we require upper bounds on the terms in the denominator and a lower bound on μ' that hold with high probability. Recall that from Lemma 2.6.8, we have

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

Also, from Lemma 2.6.6, we have that

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

Finally, by independence of Z_j,

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

which gives

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

Using these results together with (2.6.3), we have

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \le \mu-t\right) \le &\exp\left(-\min\left\{\frac{m^2n(\frac{k-1}{m}\beta_1+t-\mu)^2}{c_1(k-1)\beta_2}, \frac{mn(\frac{k-1}{m}\beta_1+t-\mu)}{c_2\beta_3}\right\}\right)\\ &+\frac{3\varepsilon}{4},\end{aligned} \tag{2.50}$$

where

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

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

Now, the first term on the right side of (2.6.3) equals $\varepsilon/4$ if

$$\mu - t = \frac{k-1}{m}\beta_1 - \max\left\{\sqrt{\frac{c_1\beta_2(k-1)}{m^2 n}}\log\frac{4}{\varepsilon}, \frac{c_2\beta_3}{mn}\log\frac{4}{\varepsilon}\right\}. \tag{2.51}$$

The expression above can be simplified under some mild assumptions on n. In particular, when $mn \geq (c_2^2/c_1)\log(4/\varepsilon)$ and $m \geq (c_2^2/c_1)\log(4n/\varepsilon)$, then (2.51) simplifies to

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

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

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

which gives us the following simplified version of (2.51) when $n \geq (c_2^2/c_1)\log(4/\varepsilon)$:

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

This completes the proof.

□

Lemma 2.6.10. *Let $Z_j, Y_{j1}, \ldots, Y_{j,k-1}$, $j \in [n]$, be independent, mean-zero random vectors in $\mathbb{R}^m$ with independent strictly subgaussian entries with variance $1/m$. Let $\mu = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n \sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2\right]$. Then, there exist absolute constants c_1 and c_2 such that for any $\varepsilon > 0$,*

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

for

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

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

Proof. The proof is similar to that of Lemma 2.6.9. We start by noting that conditioned on $\{Z_j\}_{j=1}^n$, the normalized sum $(1/n)\sum_{j=1}^n \sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2$ is subexponential with parameters v^2 and b where

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

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

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{j=1}^n\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 \geq \mu + t \middle| \{Z_j\}_{j=1}^n\right) \\
&= \Pr\left(\frac{1}{n}\sum_{j=1}^n\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2 - \mu' \geq \mu + t - \mu' \middle| \{Z_j\}_{j=1}^n\right) \\
&\leq \exp\left(-\min\left\{\frac{m^2n^2(t')^2}{c_1(k-1)\sum_{j=1}^n\|Z_j\|_2^4}, \frac{mnt'}{c_2\max_{j\in[n]}\|Z_j\|_2^2}\right\}\right) \qquad (2.52)
\end{aligned}$$

where

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

and $t' = \mu + t - \mu'$. To handle the Z_j-dependent terms in the exponent, we require high probability upper bounds on the terms in the denominator and on μ'. Proceeding as in the proof of Lemma 2.6.9, we have the following bounds on the terms in the denominator in (2.6.3):

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

and

$$\Pr\left(\max_{j\in[n]}\|Z_j\|_2^2 \le 1+\max\left\{\sqrt{\frac{c_1}{m}\log\frac{n}{\varepsilon}},\frac{c_2}{m}\log\frac{n}{\varepsilon}\right\}\right)\ge 1-\varepsilon. \tag{2.53}$$

Also, from Lemma 2.6.6,

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\|Z_j\|_2^2 \le 1+\max\left\{\sqrt{\frac{c_1}{mn}\log\frac{1}{\varepsilon}},\frac{c_2}{mn}\log\frac{1}{\varepsilon}\right\}\right)\ge 1-\varepsilon. \tag{2.54}$$

We note that although a high probability upper bound on $\max_{j\in[n]}\|Z_j\|_2^2$ implies a high probability upper bound on $(1/n)\sum_{j=1}^{n}\|Z_j\|_2^2$, we specifically use the bound in (2.54) since the deviation term has better dependence on n (which is lost in (2.53) due to a union bound step). A $\sqrt{(1/m)\log(n/\varepsilon)}$ or $(1/m)\log(n/\varepsilon)$ type dependence, on the other hand, would lead to constraints on m.

Using these results along with (2.6.3), we have

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{k-1}(Y_{jl}^\top Z_j)^2\ge\mu+t\right)\le\exp\left(-\min\left\{\frac{m^2n(\mu+t-\frac{k-1}{m}\beta_1)^2}{c_1(k-1)\beta_2},\frac{mn(\mu+t-\frac{k-1}{m}\beta_1)}{c_2\beta_3}\right\}\right)+\frac{3\varepsilon}{4}, \tag{2.55}$$

where

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

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

and

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

Simplifying as we did in Lemma 2.6.9 under the assumption that $n\ge(c_2^2/c_1)\log(4/\varepsilon)$, we

see that if

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

then the first term on the right side of (2.6.3) is less than $\varepsilon/4$, which completes the proof.

□

Lemma 2.6.11. *Let $X_1, \ldots, X_n$ be drawn i.i.d. from $\mathcal{N}(\mu_i, \sigma_i^2)$. Then, for every $t > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i^2 \le \frac{1}{n}\sum_{i=1}^n(\sigma_i^2+\mu_i^2) - t\right) \le \exp\left(\frac{-n^2t^2}{4\sum_{i=1}^n(\sigma_i^4+\sigma_i^2\mu_i^2)}\right),$$

and

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i^2 \ge \frac{1}{n}\sum_{i=1}^n(\sigma_i^2+\mu_i^2) + t\right) \le \exp\left(-\min\left\{\frac{n^2t^2}{16\sum_{i=1}^n(\sigma_i^4+\sigma_i^2\mu_i^2)}, \frac{nt}{8\max\limits_{i\in[n]}\sigma_i^2}\right\}\right).$$

Proof. The proof is similar to that of [11] for $\sigma^2 = 1$, and follows by upper bounding the MGF of a noncentral chi-squared random variable and then using the Chernoff method. We include the proof here for completeness. We will first show the left tail bound. To that end, we note that for $t > 0$ and $\lambda < 0$, the following holds for $Y \stackrel{\text{def}}{=} (1/n)\sum_{i=1}^n X_i^2$:

$$\Pr\left(Y \le \mathbb{E}\left[Y\right] - t\right) \le e^{\lambda t}\mathbb{E}\left[e^{\lambda(Y-\mathbb{E}[Y])}\right]. \tag{2.56}$$

To upper bound the MGF, first note that for $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X^2-\mathbb{E}[X^2])}\right] &= e^{-\lambda(\sigma^2+\mu^2)}\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{\lambda x^2}e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx \\ &= \frac{e^{-\lambda(\sigma^2+\mu^2)}}{\sqrt{1-2\lambda\sigma^2}}e^{\frac{\lambda\mu^2}{1-2\lambda\sigma^2}},\end{aligned}$$

for all $\lambda < 1/2\sigma^2$. Taking logarithms we have

$$\log \mathbb{E}\left[e^{\lambda(X^2-\mathbb{E}[X^2])}\right] = \frac{1}{2}\bigg(- \log(1-2\lambda\sigma^2 - 2\lambda\sigma^2)\bigg) + \frac{2\lambda^2\mu^2\sigma^2}{1-2\lambda\sigma^2} \tag{2.57}$$

$$\leq \lambda^2\sigma^4 + \frac{2\lambda^2\mu^2\sigma^2}{1-2\lambda\sigma^2}$$

$$\leq \lambda^2(\sigma^4 + 2\mu^2\sigma^2), \tag{2.58}$$

where we used $-\log(1-x) - x \leq x^2/2$ for $x < 0$ in the second step. This gives

$$\log \mathbb{E}\left[e^{\lambda(Y-\mathbb{E}[Y])}\right] \leq \frac{\lambda^2}{n^2}\sum_{i=1}^{n}(\sigma_i^4 + \sigma_i^2\mu_i^2)$$

which upon substituting into (2.56) and optimizing over $\lambda < 0$ gives $\lambda = -n^2t/(2\sum_{i=1}^{n}(\sigma_i^4 + \sigma_i^2\mu_i^2)))$ resulting in the left tail bound claimed in the lemma.

For the right tail bound, we continue from (2.57) and note that for $0 \leq \lambda \leq 1/4\sigma^2$,

$$\log \mathbb{E}\left[e^{\lambda(X^2-\mathbb{E}[X^2])}\right] \leq \frac{2\lambda^2\sigma^4}{1-2\lambda\sigma^2} + \frac{2\lambda^2\mu^2\sigma^2}{1-2\lambda\sigma^2}$$

$$\leq 4\lambda^2(\sigma^4 + \mu^2\sigma^2),$$

where in the first step we used $-\log(1-x)-x \leq x^2/2(1-x)$ for all $x \in [0,1)$. Extending as before to the normalized sum $(1/n)\sum_{i=1}^{n} X_i^2$, substituting into (2.56) and optimizing over $\lambda \in [0, 1/4\sigma^2)$, it can be seen that the minimum is attained at $\lambda = nt/(8\sum_{i=1}^{n}(\sigma_i^4 + \mu_i^2\sigma^2))$ if $t < 2\sum_{i=1}^{n}(\sigma^2+\mu^2)$, and at $\lambda = 1/(4\sum_{i=1}^{n}\sigma_i^2)$ otherwise. This gives the right tail bound claimed in the lemma. □

Lemma 2.6.12. *Let* $U_1, \ldots, U_k, V, W, Z, \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{m}I_m)$. *Then*

(i) $\mathbb{E}\left[Z^\top W\right]^2 = \frac{1}{m}$

(ii) $\mathbb{E}\left[Z^\top W\right]^4 = \frac{3}{m^2} + \frac{6}{m^3}$,

(iii) $\mathbb{E}\left[\|Z\|^4(Z^\top W)^2\right] = \frac{1}{m}\bigg(1 + \frac{6}{m} + \frac{8}{m^2}\bigg)$,

(iv) $\mathbb{E}\left[(Z^\top W)^2(Z^\top V)^2\right] = \frac{1}{m^4}\bigg(1 + \frac{2}{m}\bigg)$,

(v) $\mathbb{E}\left[\left(\sum_{i=1}^{k}(Z^\top U_i)^2\right)^2\right] = \frac{k^2}{m^2} + \frac{2k^2}{m^3} + \frac{2k}{m^2} + \frac{4k}{m^3}$.

Proof. For the proof, we use the fact that $(Z^\top W)|Z \sim \mathcal{N}(0, \frac{\|Z\|^2}{m})$.

(i)

$$\begin{aligned}\mathbb{E}\left[Z^\top W\right]^2 &= \mathbb{E}_Z\left[(Z^\top W)^2|Z\right] \\ &= \mathbb{E}\left[\frac{\|Z\|^2}{m}\right] \\ &= \frac{1}{m}.\end{aligned}$$

(ii)

$$\begin{aligned}\mathbb{E}\left[Z^\top W\right]^4 &= \mathbb{E}_Z\left[(Z^\top W)^4|Z\right] \\ &= \mathbb{E}\left[3\frac{\|Z\|^4}{m^2}\right]\end{aligned}$$

(iii)

$$\begin{aligned}\mathbb{E}\left[\|Z\|^4(Z^\top W)^2\right] &= \mathbb{E}_Z\left[\mathbb{E}\left[\|Z\|^4(Z^\top W)^2|Z\right]\right] \\ &= \mathbb{E}_Z\left[\|Z\|^4\mathbb{E}\left[(Z^\top W)^2|Z\right]\right] \\ &= \mathbb{E}\left[\frac{\|Z\|^6}{m}\right], \qquad (2.59)\end{aligned}$$

(iv)

$$\begin{aligned}\mathbb{E}\left[(Z^\top W)^2(Z^\top V)^2\right] &= \mathbb{E}_Z\left[(Z^\top W)^2(Z^\top V)^2|Z\right] \\ &= \mathbb{E}\left[\frac{\|Z\|^2}{m}\frac{\|Z\|^2}{m}\right] \\ &= \mathbb{E}\left[\frac{\|Z\|^4}{m^2}\right],\end{aligned}$$

(v)

$$\mathbb{E}\left[\left(\sum_{i=1}^{k}(U_i^\top Z)^2\right)^2\right] = \mathbb{E}_Z\left[\mathbb{E}\left[\sum_{i=1}^{k}(U_i^\top Z)^4 + \sum_{i\neq j}(U_i^\top Z)^2(U_j^\top Z)^2\middle| Z\right]\right]$$
$$= 3k\frac{\|Z\|^4}{m^2} + k(k-1)\frac{\|Z\|^4}{m^2}.$$

To complete the proof, we use Lemma 2.6.13. □

Lemma 2.6.13. *Let W and Z be m-dimensional random vectors having independent zero-mean entries with variance $1/m$ and fourth moment $3/m^2$. Then,*

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

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

$$\mathbb{E}\left[\|Z\|_2^4\right] = \sum_{i=1}^{m}\mathbb{E}\left[Z_i^4\right] + \sum_{i\neq j}\mathbb{E}\left[Z_i^2 Z_j^2\right]$$
$$= 1 + \frac{2}{m},$$

and

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

□

Lemma 2.6.14. *Let $Z, Y_1, \ldots, Y_{k-1} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{m} I_m)$ and $W = \sum_{i=1}^{k-1}(Z^\top Y_i)^2$. Then*

$$\texttt{var}(\|Z\|^4 + W) = \frac{8}{m} + \frac{12}{m^2} + \frac{50}{m^3} + \frac{10k}{m^2} + \frac{2k^2}{m^3} + \frac{16k}{m^3}. \tag{2.60}$$

Proof. Let $V = \|Z\|^4 + W$ and note that

$$\texttt{var}(V) = \mathbb{E}_Z\left[\texttt{var}(V|Z)\right] + \texttt{var}_Z(\mathbb{E}\left[V|Z\right]). \tag{2.61}$$

We start by noting that

$$\begin{aligned}\mathbb{E}\left[V|Z\right] &= \mathbb{E}\left[\|Z\|^4 + \sum_{j=1}^{k-1}(Z^\top Y_j)^2|Z\right] \\ &= \|Z\|^4 + \sum_{j=1}^{k-1}\mathbb{E}\left[(Z^\top Y_j)^2|Z\right] \\ &= \|Z\|^4 + \frac{k-1}{m}\|Z\|^2,\end{aligned} \tag{2.62}$$

which gives

$$\texttt{var}\left(\mathbb{E}\left[V|Z\right]\right) = \mathbb{E}\left[\left(\|Z\|^4 + \frac{k-1}{m}\|Z\|^2\right)^2\right] - \left(\mathbb{E}\left[\|Z\|^4 + \frac{k-1}{m}\|Z\|^2\right]\right)^2. \tag{2.63}$$

Consider the first term:

$$\mathbb{E}\left[\left(\|Z\|^4 + \frac{k-1}{m}\|Z\|^2\right)^2\right] = \mathbb{E}\left[\|Z\|^8\right] + \frac{2(k-1)}{m}\mathbb{E}\left[\|Z\|^6\right] + \left(\frac{k-1}{m}\right)^2\mathbb{E}\left[\|Z\|^4\right]. \tag{2.64}$$

Using Lemma 2.6.13 we get

$$\begin{aligned}\mathbb{E}\left[\left(\|Z\|^4 + \frac{k-1}{m}\|Z\|^2\right)^2\right] =& 1 + \frac{12}{m} + \frac{44}{m^2} + \frac{48}{m^3} + \frac{2(k-1)}{m}\left(1 + \frac{6}{m} + \frac{8}{m^2}\right) \\ &+ \left(\frac{k-1}{m}\right)^2\left(1 + \frac{2}{m}\right).\end{aligned} \tag{2.65}$$

Now consider the second term in (2.63):

$$\begin{aligned}\mathbb{E}\left[\|Z\|^4 + \frac{k-1}{m}\|Z\|^2\right] &= \left(1 + \frac{2}{m}\right) + \frac{k-1}{m} \\ &= 1 + \frac{k+1}{m}.\end{aligned} \tag{2.66}$$

Combining the above terms, (2.63) becomes

$$\begin{aligned}\mathtt{var}_Z\mathbb{E}\left[V|Z\right] &= 1+\frac{12}{m}+\frac{26}{m^2}+\frac{66}{m^3}+\frac{2(k-1)}{m}\left(1+\frac{6}{m}+\frac{8}{m^2}\right)+\left(\frac{k-1}{m}\right)^2\left(1+\frac{2}{m}\right)\\ &\quad-\left(1+\frac{k+1}{m}\right)^2\\ &= \frac{8}{m}+\frac{32}{m^2}+\frac{34}{m^3}+\frac{8k}{m^2}+\frac{2k^2}{m^3}+\frac{12k}{m^3}. \end{aligned} \tag{2.67}$$

This gives us one component of the variance in (2.61). We now compute the other component, $\mathbb{E}\left[Z\right]\mathtt{var}(V|Z)$. Recall that

$$V = \|Z\|^4+\sum_{j=1}^{k-1}(Z^\top Y_j)^2. \tag{2.68}$$

Note that

$$\begin{aligned}\mathtt{var}(V|Z) &= \mathtt{var}\left(\|Z\|^4+\sum_{j=1}^{k-1}(Z^\top Y_j)^2|Z\right)\\ &= \sum_{j=1}^{k-1}\mathtt{var}\left((Z^\top Y_j)^2|Z\right)\\ &= \sum_{j=1}^{k-1}\left(\frac{3}{m^2}\|Z\|^4-\frac{1}{m^2}\|Z\|^4\right)\\ &= \frac{2(k-1)}{m^2}\|Z\|^4.\end{aligned}$$

where we used the same argument as in Lemma 2.6.12 to get the third step. And so, this gives

$$\begin{aligned}\mathbb{E}_Z\left[\mathtt{var}(V|Z)\right] &= \frac{2(k-1)}{m^2}\left(1+\frac{2}{m}\right)\\ &= \frac{2(k-1)}{m^2}+\frac{4(k-1)}{m^3}.\end{aligned} \tag{2.69}$$

Combining (2.61), (2.67) and (2.69), we get

$$\texttt{var}(\|Z\|_2^4 + W) = \frac{8}{m} + \frac{12}{m^2} + \frac{50}{m^3} + \frac{10k}{m^2} + \frac{2k^2}{m^3} + \frac{16k}{m^3}. \tag{2.70}$$

□

Lemma 2.6.15. *Let $X \sim \mathcal{N}(0, I_d)$ and $B \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Then*

$$\texttt{var}(X^\top BX) = 2\ \texttt{Tr}(B^2). \tag{2.71}$$

Proof. We start by noting that $\mathbb{E}\left[X^\top BX\right] = \texttt{Tr}(B)$, which gives

$$\begin{aligned}
\texttt{var}(X^\top BX) &= \mathbb{E}\left[(X^\top BX - \texttt{Tr}(B))^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^d B_{ii}(X_i^2 - 1) + \sum_{i\neq j} B_{ij}X_iX_j\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^d B_{ii}(X_i^2 - 1)\right)^2\right] + \mathbb{E}\left[\left(\sum_{i\neq j} B_{ij}X_iX_j\right)^2\right] \\
&\quad + 2\mathbb{E}\left[\sum_{i=1}^d B_{ii}(X_i^2 - 1)\sum_{i\neq j} B_{ij}X_iX_j\right].
\end{aligned}$$

We evaluate each of the three terms separately. For the first term,

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^d B_{ii}(X_i^2 - 1)\right)^2\right] &= \mathbb{E}\left[\sum_{i=1}^d B_{ii}^2(X_i^2 - 1)^2 + \sum_{i\neq j} B_{ii}B_{jj}(X_i^2 - 1)(X_j^2 - 1)\right] \\
&= 2\sum_{i=1}^d B_{ii}^2,
\end{aligned}$$

where we used $\mathbb{E}\left[X_i^2\right] = 1$ and $\mathbb{E}\left[X_i^4\right] = 3$. Similar calculations for the second and third terms give

$$\mathbb{E}\left[\left(\sum_{i\neq j} B_{ij}X_iX_j\right)^2\right] = 4\mathbb{E}\left[\sum_{i<j} B_{ij}^2X_i^2X_j^2 + \sum_{i,j,k} B_{ij}B_{ik}X_i^2X_jX_k + \sum_{i,j,k,l} B_{ij}B_{kl}X_iX_jX_kX_l\right]$$

$$= 4 \sum_{i<j} B_{ij}^2,$$

and

$$\mathbb{E}\left[\sum_{i=1}^{d} B_{ii}(X_i^2 - 1) \sum_{i \neq j} B_{ij} X_i X_j\right] = 0.$$

Combining everything, we get

$$\begin{aligned} \texttt{var}(X^\top B X) &= 2 \sum_{i=1}^{d} B_{ii}^2 + 4 \sum_{i<j} B_{ij}^2 \\ &= 2 \sum_{i,j=1}^{d} B_{ij}^2 \\ &= 2 \ \texttt{Tr}(B^2), \end{aligned}$$

where the last step uses the fact that B is symmetric.

□

Chapter 3

Recovering a Single Support: Lower bound

In this chapter, we derive a lower bound on the sample complexity of common support recovery that matches the upper bound obtained in the previous chapter, thus determining the optimal tradeoff between m and n in the $m < k$ regime. Our result shows a phase transition that occurs at $k/m = 1$ for the problem of support recovery when there is a single unknown support. In particular, the dependence of the sample complexity on k/m undergoes a sharp change from linear to quadratic as we move from the $k/m \leq 1$ regime to the $k/m > 1$ regime.

3.1 Lower bound

Theorem 3.1.1. *For $1 \leq m < k/2$, $1 \leq k \leq d-1$, and $\sigma^2 = 0$, the sample complexity of support recovery satisfies*

$$n^*_{\mathsf{C},\mathsf{avg}}(m,k,d) = \Omega\left(\frac{k^2}{m^2}\log(k(d-k))\right).$$

Proof. We work with the Gaussian setting, with the samples and measurement matrices

The work in this chapter is based on [60], [58].

satisfying Assumptions 2.3.3 and 2.3.4. Denote by $\mathcal{S}_0$ the set $\{1, \ldots, k\}$ and by $\mathcal{S}_{i,j}$, $1 \leq i \leq k < j \leq d$, the set obtained by replacing the element i in $\mathcal{S}_0$ with j from $\mathcal{S}_0^c$. Let U be distributed uniformly over the pairs $\{(i,j) : 1 \leq i \leq k, k+1 \leq j \leq d\}$. The unknown support is set to be $\mathcal{S}_U$; the random variables X_1^n and linear measurements $Y_i = \Phi_i X_i$ are generated as before.

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

$$\begin{aligned}\Pr\left(\hat{U} \neq U\right) &\geq 1 - \frac{I(Y_1^n; U) + 1}{\log(k(d-k))} \\ &\geq 1 - \frac{\max_u D(\mathrm{P}_{Y^n|\mathcal{S}_u} \| \mathrm{P}_{Y^n|\mathcal{S}_0}) + 1}{\log(k(d-k))},\end{aligned}$$

where $\mathrm{P}_{Y^n|\mathcal{S}}$ denotes the distribution of the measurements when the support of λ is $\mathcal{S}$ (a proof for the second inequality can be found in [23, Theorem 21]). Note that $\mathrm{P}_{Y^n|\mathcal{S}} = \prod_{i=1}^n \mathrm{P}_{Y_i|\mathcal{S}}$ with each $\mathrm{P}_{Y_i|\mathcal{S}}$ having the same distribution which we denote by $\mathrm{P}_{Y|\mathcal{S}}$. Thus, $D(\mathrm{P}_{Y^n|\mathcal{S}_u} \| \mathrm{P}_{Y^n|\mathcal{S}_0}) = nD(\mathrm{P}_{Y|\mathcal{S}_u} \| \mathrm{P}_{Y|\mathcal{S}_0})$.

Next, we bound $D(\mathrm{P}_{Y|\mathcal{S}_u} \| \mathrm{P}_{Y|\mathcal{S}_0})$. Denote by $\Phi_{\mathcal{S}}$ the $m \times k$ submatrix of Φ obtained by restricting to the columns in $\mathcal{S}$ and by $A_{\mathcal{S}}$ the Gram matrix $\Phi_{\mathcal{S}}\Phi_{\mathcal{S}}^\top$ of $\Phi_{\mathcal{S}}$. Further, let $\nu_1 \geq \ldots \geq \nu_m > 0$ and $\nu'_1 \geq \ldots \geq \nu'_m > 0$ be the respective eigenvalues of $A_{\mathcal{S}_u}$ and $A_{\mathcal{S}_0}$. Note that $\nu_m > 0$ and $\nu'_m > 0$ hold with probability 1 since $m \leq k$.

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

$$\begin{aligned}D\big(\mathrm{P}_{Y|\mathcal{S}_u,\Phi} \| \mathrm{P}_{Y|\mathcal{S}_0,\Phi}\big) &= \frac{1}{2}\left(\log\frac{|A_{\mathcal{S}_0}|}{|A_{\mathcal{S}_u}|} + \mathtt{Tr}(A_{\mathcal{S}_0}^{-1}A_{\mathcal{S}_u}) - m\right) \\ &\leq \frac{1}{2}\sum_{i=1}^m \left(\log\frac{\nu'_i}{\nu_i} - \left(1 - \frac{\nu_i}{\nu'_i}\right)\right) \\ &\leq \frac{1}{2}\sum_{i=1}^m \frac{(\nu_i - \nu'_i)^2}{\nu_i\nu'_i},\end{aligned}$$

where in the first inequality holds by Lemma 3.4.2 and the second inequality holds since $\log x + (1-x)/x \leq (x-1)^2/x$ for all $x > 0$. Using convexity of the KL divergence, we get

$$\begin{aligned} D\big(\mathrm{P}_{Y|\mathcal{S}_u} \| \mathrm{P}_{Y|\mathcal{S}_0}\big) &\leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \frac{(\nu_i - \nu_i')^2}{\nu_i \nu_i'}\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \frac{(\nu_i - \nu_i')^2}{\nu_m \nu_m'}\right]. \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{m} \frac{(\nu_i - \nu_i')^2}{\nu_m \nu_m'}\right] &\leq \sqrt{\mathbb{E}\left[\frac{1}{\nu_m^2 \nu_m'^2}\right]} \sqrt{\mathbb{E}\left[\Big(\sum_{i=1}^{m} (\nu_i - \nu_i')^2\Big)^2\right]} \\ &\leq \sqrt{\mathbb{E}\left[\frac{1}{\nu_m^4}\right]} \sqrt{\mathbb{E}\left[\Big(\sum_{i=1}^{m} (\nu_i - \nu_i')^2\Big)^2\right]}, \end{aligned}$$

where in the second inequality we also used the fact that a_i and b_i are identically distributed. The Hoffman-Wielandt inequality[1] [33] can be used to handle the second term on the right-side. In particular, we have $\sum_{i=1}^{m} (\nu_i - \nu_i')^2 \leq \|A_{\mathcal{S}_0} - A_{\mathcal{S}_u}\|_F^2$ where the right-side coincides with $\|\Phi_1 \Phi_1^\top - \Phi_{k+1} \Phi_{k+1}^\top\|_F^2$ since $u = (1, k+1)$. Using the triangle inequality for Frobenius norm and noting that $\|\Phi_i \Phi_i^\top\|_F$ equals $\|\Phi_i\|_2^2$ for a vector Φ_i, we get

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

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

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

Using the expression for the fourth moment of a chi-squared random variable gives us

$$\mathbb{E}\left[\sum_{i=1}^{m}\frac{(\nu_i-\nu_i')^2}{\nu_m\nu_m'}\right] \leq \sqrt{\mathbb{E}\left[\frac{1}{\nu_m^4}\right]}\sqrt{\frac{1}{m^4}\frac{(m+3)!}{(m-1)!}}$$
$$\leq c'\sqrt{\mathbb{E}\left[\frac{1}{\nu_m^4}\right]}$$

where c' is an absolute constant.

It only remains to bound $\mathbb{E}\left[1/\nu_m^4\right]$, where ν_m is the minimum eigenvalue of the $(m\times m)$ Wishart matrix $A_{\mathcal{S}_u}$. Using Lemma 3.4.1, we can obtain

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

By combining all the steps above, we get

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

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

□

Remark 3.1.2. *We note that our lower bound proof requires some separation between k and m; namely, it requires $k/m > \gamma$ for some $\gamma > 1$. While the lower bound of $n = \Omega((k/m)\log(d/k))$ from previous work [55] continues to hold for $m < k$, it is not clear if a tighter lower bound on sample complexity in the regime $1 < k/m \leq \gamma$ can be obtained. Such a separation between k and m is, however, not required when deriving the upper bound.*

3.1.1 Extension to nonbinary variances

Theorem 3.1.3. *When $\sigma^2 = 0$, we have for an absolute constant c the lower bound*

$$n^*_{\mathtt{C},\mathtt{avg}}(m,k,d) \geq c\frac{\lambda^2_{max}}{\lambda^2_{min}}\frac{k^2}{m^2}\log(d-k+1).$$

We assume that the unknown λ is uniformly distributed over the set $\{\lambda^{(0)}, \lambda^{(1)}, \ldots, \lambda^{(d-k)}\}$, with $\lambda^{(i)} \in \mathbb{R}^d$. The jth entry of $\lambda^{(i)}$, denoted $\lambda^{(i)}_j$, is given by

$$\lambda^{(i)}_j = \begin{cases} \lambda_{\max}, & \text{if } j \in [k-1], \\ \lambda_{\min}, & \text{if } j = k+i, \\ 0, & \text{otherwise,} \end{cases}$$

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

Our goal is to characterize the KL divergence between distributions on the measurements arising from two different λs in the set we described above, one of which we fix as $\lambda^{(0)}$. Computing this divergence as before, we see that

$$D\big(\mathrm{P}_{Y|\lambda}\|\mathrm{P}_{Y|\lambda^{(0)}}\big) \leq \frac{1}{2}\,\mathbb{E}\left[\sum_{i=1}^{m}\frac{(\kappa_i-\kappa'_i)^2}{\kappa_m\kappa'_m}\right], \tag{3.1}$$

where $\{\kappa_i\}_{i=1}^m$ and $\{\kappa'_i\}_{i=1}^m$ denote the eigenvalues of $A_\lambda \stackrel{\text{def}}{=} \Phi K_\lambda \Phi^\top$ and $A_{\lambda^{(0)}} \stackrel{\text{def}}{=} \Phi K_{\lambda^{(0)}}\Phi^\top$ respectively. Noting that $\sum_{i=1}^m(\kappa_i-\kappa'_i)^2 \leq \|A_\lambda - A_{\lambda^{(0)}}\|_F^2 = \lambda^2_{\min}\|\Phi_1\Phi_1^\top - \Phi_{k+1}\Phi_{k+1}^\top\|_F^2$, an application of the Hoffman-Wielandt inequality yields

$$\mathbb{E}\left[\sum_{i=1}^{m}\frac{(\kappa_i-\kappa'_i)^2}{\kappa_m\kappa'_m}\right] \leq c\lambda^2_{\min}\sqrt{\mathbb{E}\left[\frac{1}{(\kappa_m)^4}\right]}. \tag{3.2}$$

Recall that from Lemma 3.4.1, we have a bound on the fourth moment of the smallest eigenvalue ν_m of $A_S = \Phi_S\Phi_S$ for $S \subseteq [d]$. We now try to relate κ_m and ν_m. We start by

noting that

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

where $A \succcurlyeq B$ if $A - B$ is a positive semi-definite matrix. The above inequality in turn gives a relation between the eigenvalues of A_λ and those of $\lambda_{\max} \sum_{i=1}^{k-1} \Phi_i \Phi_i^\top$. In particular, for the minimum eigenvalue, we have $\kappa_m \geq \lambda_{\max} \nu_m$. Combining this fact with the inequalities in (3.1) and (3.2), and using Lemma 3.4.1, we get

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

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

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

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

3.2 A phase transition for support recovery

The lower bound from Theorem 3.1.1 for Gaussian inputs implies a lower bound for worst-case inputs as well, since an instantiation in the Gaussian case can be thought of as a deterministic input. In particular, we have $n^*_{\mathtt{C},\mathtt{w}}(m,k,d) \geq n^*_{\mathtt{C},\mathtt{avg}}(m,k,d)$. In fact, the dependence of our upper bounds derived in Chapter 2 and the lower bound from previous section on the problem dimensions (m,k,d) coincides. We will use $n^*(m,k,d)$ to denote this common scaling. Combining Theorems 2.3.1 and 3.1.1, we obtain the following tight

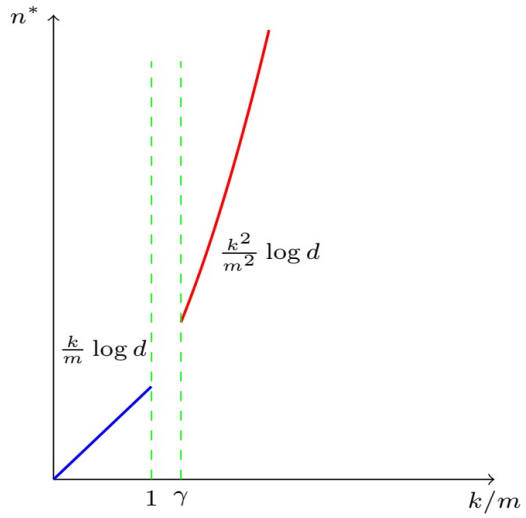


Figure 3.1: Sample complexity of support recovery as a function of k/m.

characterization of $n^*(m, k, d)$.

Theorem 3.2.1 (Characterization of sample complexity)**.** *For* $(\log k)^2 \leq m < k/2$ *and* $1 \leq k \leq d/2$*, the sample complexity of common support recovery is given by*

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

Remark 3.2.2. *We expect the scaling in Theorem 3.2.1 to hold good even when* $m < (\log k)^2$*. In fact, our lower bound result continues to hold for* $m = 1$*. The current upper bound proof, however, requires* $m \geq (\log k)^2$*.*

Remark 3.2.3. *As long as the noise variance is sufficiently small, i.e.,* $\sigma^2 < k/m$*, our estimator is sample-optimal and achieves the same scaling as the lower bound.*

In summary, our results settle the question of tradeoff between m and n in the $m < k$ regime, and show that there exists a phase transition for the sample complexity of this problem at $k/m = 1$ as depicted in Figure 3.2. Roughly, around this point, the sample complexity for support recovery undergoes a change from being linear in the ratio k/m to being quadratic in k/m (up to a factor of $\log d$).

3.3 Discussion

We showed a phase transition for the problem of support recovery from multiple samples. While the closed form estimator that we analyzed here is sample-optimal, it would be interesting to design other estimators that can work in the measurement-constrained regime without knowledge of the support size, and for which guarantees can be obtained with worst-case inputs. Finally, extending the lower bound on n^* to include the $1 < k/m \leq \gamma$ regime for $\gamma > 1$ would provide a better understanding of the problem.

3.4 Remaining Proofs

Lemma 3.4.1. *Let $\Phi \in \mathbb{R}^{m\times k}$ with independent $\mathcal{N}(0,1)$ entries and let $A = \Phi\Phi^\top$. If Z denotes the minimum eigenvalue of A, then for $k - m > 7$,*

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

Proof. Since Z is a nonnegative random variable, we have for $\theta > 0$,

$$\begin{aligned}\mathbb{E}\left[Z^{-4}\right] &\leq \theta + \int_\theta^\infty \Pr\left(Z \leq u^{-\frac{1}{4}}\right) du \\ &\leq \theta + \frac{8}{k^4}\int_0^{\frac{\theta^{-\frac{1}{8}}}{\sqrt{k}}} \Pr\left(Z \leq k\varepsilon^2\right)\frac{1}{\varepsilon^9} d\varepsilon,\end{aligned}$$

where we used $u^{-\frac{1}{4}} = k\varepsilon^2$. The density of the smallest eigenvalue of a Wishart matrix with parameters k and m (A in this case) is known in closed form [22, Lemma 4.1], which we restate here:

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

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

get

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

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

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

□

Lemma 3.4.2. *Let $A, B \in \mathbb{R}^{m\times m}$ be symmetric, positive definite matrices and let $a_1 \geq \cdots \geq a_m$ and $b_1 \geq \cdots \geq b_m$ denote their respective ordered eigenvalues. Then,*

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

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

$$\mathtt{Tr}(AB) = \sum_{i=1}^{m} \gamma_i \leq \sum_{i=1}^{m} |\gamma_i| \leq \sum_{i=1}^{m} s_i, \tag{3.3}$$

where the last inequality follows from [34] [Theorem 3.3.13]. The next step is to relate the sum of the singular values of AB to the eigenvalues of A and B. We use the following two results from [34] [Theorem 3.3.4, Corollary 3.3.10]:

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

$$\prod_{i=1}^{m} s_i \leq \prod_{i=1}^{m} a_i b_i; \tag{3.4}$$

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

$$\prod_{i=1}^{m} \alpha_i \leq \prod_{i=1}^{m} \beta_i, \tag{3.5}$$

then

$$\sum_{i=1}^{m} \alpha_i \leq \sum_{i=1}^{m} \beta_i. \tag{3.6}$$

From the results above, we have that

$$\sum_{i=1}^{m} s_i \leq \sum_{i=1}^{m} a_i b_i, \tag{3.7}$$

which together with (3.3) gives the result. □

Chapter 4

Recovering Multiple Supports

In this chapter, we study the problem of multiple support recovery, where 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.

4.1 Introduction

In the problem of *multiple support recovery*, 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,* 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

The work in this chapter is based on [61].

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 which has been widely studied [72], [26], [88], [48], [55], [60]. It is also related to the study of sparse random effects in mixed linear models [6, 8]. 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 [46, 80] 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 [81], [4]. 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 chapter, we will show that one can operate in the measurement-constrained regime of $m < k$, when a sufficiently large number of samples is available.

4.1.1 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 [81] 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 [72], [26], [55].

On the other hand, in the $m < k$ regime, we know from the previous chapters and from [58], [60] 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, as we saw, the lower bound from Chapter 3 applies to the worst-case setting as well, showing that while k overall measurements suffice when m exceeds k, at least (roughly) k^2/m measurements are required when $m < k$.

In [51], 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 [86], [57], [5] that use hierarchical Bayesian models and 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 [89], [41], [43], [19], [5], [50]. 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 as we saw in Chapter 2 and requires $\Omega(k^2 \log d)$ samples to

recover the common support.

4.1.2 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 closed-form estimator from Chapter 2. We then construct an affinity matrix using the variance estimates from the first step 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 [45], [49], [32], [1], 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 different 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 4.2.1. In similar spirit to Chapter 2, 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 4.5. For the case of two supports, we provide an analysis for intersecting supports.

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 4.3, and analyze its performance in Section 4.4. We provide experimental results in Section 4.5. The technical results required for the proofs in Section 4.4 are available in the appendices.

4.2 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[1]

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

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

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

$$\mathrm{P}_{\mathcal{S}_1,\ldots,\mathcal{S}_\ell} = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathrm{P}^{(i)}, \tag{4.1}$$

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 Section 4.9 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 4.2.1. *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 4.2.2. *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 4.2.1. *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*

$$\mathrm{P}_{\mathcal{S}_1,\ldots,\mathcal{S}_\ell}\left(\exists\, \sigma \in \mathcal{G}_\ell \ s.t. \ \sum_{i=1}^{\ell} \left|\mathcal{S}_i \Delta \hat{\mathcal{S}}_{\sigma(i)}\right| < k\varepsilon\ell^2\right) \geq 1 - \delta, \tag{4.2}$$

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 an (n, ε, δ)-estimator using a small number of samples of n. For fixed $m, k, d, \ell, \varepsilon$, 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 $n^*_{\mathtt{M,avg}}(m, k, d, \ell, \varepsilon, \delta)$. In our main result stated below, we provide an upper bound on this quantity.

Theorem 4.2.1. *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 4.2.1 and 4.2.2, the sample complexity of multiple support recovery satisfies*

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

Remark 4.2.2. *For values of ε lower than $1/k\ell$, the result from Theorem 4.2.1 continues to hold with ε set to $1/k\ell$. This is because $\varepsilon = 1/k\ell$ 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 4.4.3.

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 Chapter 2 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.

4.3 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 Chapter 2. Following this, we use a spectral clustering based approach to recover the individual supports. We provide more details in the next two subsections.

4.3.1 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 Chapter 2 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 closed-form estimator and form the statistic $\tilde{\lambda} \in \mathbb{R}^d$ as follows. First, define vectors $a'_1, \ldots, a'_n$ with entries

$$a'_{ji} \stackrel{\text{def}}{=} (\Phi_{ji}^\top Y_j)^2, \quad i \in [d]. \tag{4.3}$$

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

$$\tilde{\lambda} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} a'_j. \tag{4.4}$$

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 Chapter 2, 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

$$\hat{\mathcal{S}}_{\text{un}} = \{(1), \ldots, (k\ell)\}, \tag{4.5}$$

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 [92].

4.3.2 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 (4.3) 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 $\hat{\mathcal{S}}_{\text{un}} = [k\ell]$.[2]

Our approach is the following: we construct a $k\ell \times k\ell$ *affinity matrix* T and perform

[2]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 2.

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

Figure 4.1: Block structure of the expected affinity matrix when $\ell = 2$ and the supports are disjoint, under appropriate permutation of rows and columns.

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

$$T_{uv} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} a_{ju}a_{jv}, \quad (u, v) \in [k\ell] \times [k\ell]. \tag{4.6}$$

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 4.1 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 [49, 65], 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 2.

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 operator norm, which for symmetric matrices is the largest eigenvalue in magnitude. The key challenge, therefore, is to control $\|T - \mathbb{E}[T]\|_{op}$.

Unlike typical settings, the entries of the affinity matrix T in our case 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 (4.6) is unbounded, so deriving a tail bound for the sum requires a more careful tail splitting method (see, for example, [73, Exercise 2.1.7]), and leads to measurement matrix dependent quantities that are difficult to handle. As we will see shortly, the matrix T can be expressed as a sum of rank one matrices, and so one approach could be to apply techniques from matrix concentration to obtain tail bounds for $\|T - \mathbb{E}[T]\|_{op}$. These techniques, however, either require the summands to be bounded almost surely in spectral norm or to have subexponential-type moments [76, Theorem 6.1, 6.2], neither of which is true in our case.

To circumvent this difficulty, we turn to a beautiful result by Rudelson [66], 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 (4.6) can equivalently be represented as

Algorithm 2: Multiple support recovery: General case

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

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

$T = (1/n)\sum_{i=1}^{n} a_i a_i^\top$. 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. Although Rudelson's result characterizes the expected operator norm, it has since been extended to handle higher moments and tails, see [67], [52] for more details.

4.4 Analysis of the estimator

We will first analyze the performance of the union recovery step. Then, conditioned on the union being exactly recovered, we analyze the second step of our estimator.

4.4.1 Recovering the union: Analysis

Our analysis of the probability of exactly recovering $\mathcal{S}_{\text{un}}$ using the estimator in (4.5) follows the approach in Chapter 2. 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 (4.1), 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 Chapter 2. We prove this result for the two component mixture; it can be extended easily to the general case.

Lemma 4.4.1. *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

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

$$\leq e^{\frac{\theta^2 c^2}{2}},$$

where $c = \max\{a, b\}$. □

Thus, the samples $X_1, X_2, \ldots, X_n$ 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 Chapter 2 imply that we can recover $\mathcal{S}_{\text{un}}$ from the variance estimate (4.4) by retaining the $k\ell$ largest entries. In particular, a direct application of [60, Theorem 3] with support size set to $k\ell$, gives us the following result.

Theorem 4.4.2. *Let $\hat{\mathcal{S}}_{un}$ described in* (4.5) *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 closed-form estimator 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.

4.4.2 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 4.4.3 (Block structure of $\mathbb{E}[T]$)**.** *Under Assumptions 4.2.1 and 4.2.2, for the matrix $T \in \mathbb{R}^{k\ell \times k\ell}$ in (4.6), $\mathbb{E}[T]$ 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 4.4.3 appears in Section 4.8.5 and involves computing the expected values of random variables that contain higher order terms in Φ_i and X_i. Before we proceed, we note the following extension of the "median trick" (see, for example, [18]) 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 4.4.4 (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 Section 4.8.1.

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 4.4.5. *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 4.4.5 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 4.4.5.

Lemma 4.4.6. *Under Assumptions 4.2.1 and 4.2.2, 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 4.4.7. *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 4.4.6 is provided in Section 4.8.6 and the proof of Lemma 4.4.7 appears in Section 4.8.2.

We close this section with the proof of Theorem 4.4.5.

Proof of Theorem 4.4.5. 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 2 labels more than $\varepsilon k\ell$ coordinates incorrectly by $\mathcal{E}$. An upper bound on $\Pr(\mathcal{E})$ would imply an upper bound on the probability of the error event implied by (4.2). The per support errors across the ℓ labels can have significant overlap or even be equal, so the criterion in (4.2) is a good indicator of the number of misclustered coordinates determined by $\mathcal{E}$. Additionally, it satisfies the triangle inequality, a property we will use later in proving Lemma 4.4.4.

The following result relates the error probability to a perturbation bound.

Lemma 4.4.8 (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,*

$$\Pr(\mathcal{E}) \leq \Pr\left(\|\hat{V} - VO\|_F \geq \frac{1}{2}\sqrt{\frac{\varepsilon\ell}{2}}\right), \tag{4.7}$$

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 [65] and requires us to use some properties of V, which we note in the lemma below.

Lemma 4.4.9 (Properties of V)**.** *For* $1 \leq i \leq k\ell$*, denote by* v^i *the* i*th 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 4.4.9 in Section 4.8.3.

Proof of Lemma 4.4.8. We begin by observing that it suffices to show that

$$\Pr(\mathcal{E}) \leq \Pr\left(\|C - VO\|_F \geq \sqrt{\frac{\varepsilon\ell}{2}}\right), \tag{4.8}$$

where C is the matrix found in Step 6 of Algorithm 2 and is random since $\hat{V}$ is random. Indeed, by Lemma 4.4.9, 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

$$\begin{aligned}
\|C - VO\|_F &\leq \|C - \hat{V}\|_F + \|VO - \hat{V}\|_F && (4.9)\\
&= \min_{U\in\mathcal{U}_\ell} \|U - \hat{V}\|_F + \|VO - \hat{V}\|_F && (4.10)\\
&\leq 2\|VO - \hat{V}\|_F, && (4.11)
\end{aligned}$$

where the final bound holds since VO belongs to $\mathcal{U}_\ell$. Thus, (4.8) will imply (4.7). 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 (4.8). To that end, we define

$$\mathcal{I} \stackrel{\text{def}}{=} \{i \in [k\ell] : \|v^i O - c^i\|_2 < 1/\sqrt{2k}\}, \tag{4.12}$$

where v^i and c^i are the ith row of V and C, respectively. Our claim is that Algorithm 2 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

$$\begin{aligned} \|v^i O - v^j O\|_2 &\leq \|v^i O - c^j\|_2 + \|v^j O - c^j\|_2 && (4.13)\\ &\leq \|v^i O - c^i\|_2 + \|c^i - c^j\|_2 + \|v^j O - c^j\|_2 && (4.14)\\ &< \sqrt{\frac{2}{k}} + \|c^i - c^j\|_2. && (4.15) \end{aligned}$$

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 4.4.9 implies that $v^i O = v^j O$. Therefore, when the labels given by the algorithm for coordinates i and j coincide (this happens only when $c^i = c^j$), then $v^i O = v^j O$. But then, by the first property in Lemma 4.4.9, 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 $|\mathcal{I}^c| \leq \varepsilon k\ell < k$, then at least one element from each part $\mathcal{S}_1, \ldots, \mathcal{S}_\ell$ 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 $|\mathcal{I}^c| \leq \varepsilon k\ell$. Then, clearly, in this case the error event $\mathcal{E}$ does not hold. It follows from the definition of $\mathcal{I}$ that

$$\begin{aligned}
\Pr(\mathcal{E}) &\leq \Pr(|\mathcal{I}^c| > \varepsilon k\ell) && (4.16)\\
&\leq \Pr\left(\left|\left\{i : \|c^i - v^i O\|_2 \geq \frac{1}{\sqrt{2k}}\right\}\right| > \varepsilon k\ell\right) && (4.17)\\
&\leq \Pr\left(\|C - VO\|_F^2 > \frac{\varepsilon\ell}{2}\right), && (4.18)
\end{aligned}$$

where in the final step we used the fact that the second step implies $\|C - VO\|_F^2 = \sum_{i=1}^{k\ell} \|c^i - v^i O\|_2^2 \geq \varepsilon k\ell/2k$. This completes the proof of (4.8).

□

Step 2: Controlling the perturbation.

In view of Lemma 4.4.8, 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 4.4.10 (Perturbation of eigenspace)**.** *[90] Let A and $\hat{A}$ be $d \times d$ symmetric matrices with eigenvalues $\nu_1 \geq \cdots \geq \nu_d$ and $\hat{\nu}_1 \geq \cdots \geq \hat{\nu}_d$, 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*

$$\|\hat{V} - VO\|_F \leq 2\sqrt{2}\, \frac{\min\{\sqrt{\ell}\|\hat{A} - A\|_{op}, \|\hat{A} - A\|_F\}}{\nu_\ell - \nu_{\ell+1}}. \tag{4.19}$$

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

$$\|\hat{V} - VO\|_F \leq \frac{2\sqrt{2}}{\Delta_\ell} \min\{\sqrt{\ell}\|T - \mathbb{E}[T]\|_{op}, \|T - \mathbb{E}[T]\|_F\}, \tag{4.20}$$

where $\Delta_\ell \stackrel{\text{def}}{=} \nu_\ell - \nu_{\ell+1}$. Combining this bound with our earlier bound from Lemma 4.4.8,

we get

$$\Pr(\mathcal{E}) \leq \Pr\left(\|T - \mathbb{E}[T]\|_{op} \geq \frac{\Delta_\ell \sqrt{\varepsilon}}{8} \right) \tag{4.21}$$

$$\leq \frac{8}{\Delta_\ell \sqrt{\varepsilon}} \cdot \mathbb{E}\left[\|T - \mathbb{E}[T]\|_{op} \right], \tag{4.22}$$

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 [66]. As pointed out earlier, the original bound in [66] 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 Section 4.8.4.

Theorem 4.4.11 (Extension of a result in [66])**.** *Let $Z \in \mathbb{R}^N$ be a random vector such that $A = \mathbb{E}\left[ZZ^\top\right]$. Let $Z_1, \ldots, Z_n$ be independent copies of Z. Then, there exists an absolute constant c such that*

$$\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top - A \right\|_{op} \right] \leq \frac{1}{2} \left(\alpha^2 + \alpha \sqrt{\alpha^2 + 4\|A\|_{op}} \right), \tag{4.23}$$

where

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

Using this bound in (4.22) with $N = k\ell$, we obtain

$$\Pr(\mathcal{E}) \leq \frac{4}{\Delta_\ell \sqrt{\varepsilon}} \left(\alpha^2 + \alpha \sqrt{\alpha^2 + 4\|\mathbb{E}[T]\|_{op}} \right). \tag{4.24}$$

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

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

□

In the next section, we combine the results from Theorems 4.4.2 and 4.4.5 to show the sample complexity bound of Theorem 4.2.1.

4.4.3 Proof of Theorem 4.2.1

The proof of Theorem 4.2.1 now follows by combining guarantees for the union recovery step from Theorem 4.4.2 and the clustering step from Theorem 4.4.5.

We begin by applying Theorem 4.4.2 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 4.4.2 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 4.4.6 and Lemma 4.4.7 into Theorem 4.4.5, we see that for ε-approximate recovery of the supports it suffices to have

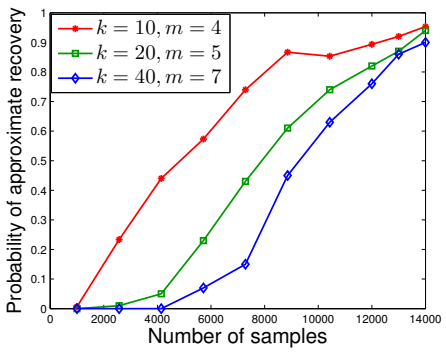
$$\begin{aligned} n &\geq \frac{c}{\varepsilon}\lambda_0^2 \frac{k^3\ell}{m^2}\frac{\ell^2}{\lambda_0^4 k^2} \cdot n^{\frac{2}{\log k}} \cdot \left(\lambda_0 \frac{k\sqrt{k}\sqrt{\ell}}{m}(\log k)^2\right)^2 \cdot \log(k\ell) \\ &= \frac{c}{\varepsilon}\frac{k^4\ell^4}{m^4} n^{\frac{2}{\log k}} (\log k)^4 \log(k\ell). \end{aligned} \tag{4.25}$$

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.

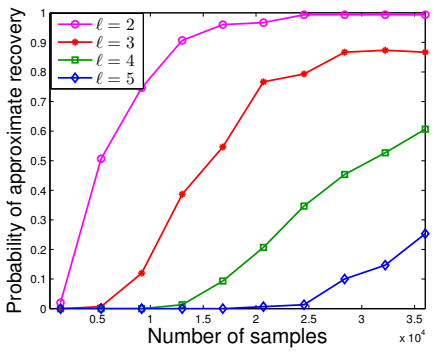
4.5 Simulations

4.5.1 Synthetic data

In this subsection, we evaluate the performance of Algorithm 2 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.



(a) $d = 100$, $\varepsilon = 0.2$, $\ell = 2$.



(b) $d = 100$, $\varepsilon = 0.2$, $m = 4$, $k = 10$.

Figure 4.2: Probability of approximate support recovery with (a) varying k/m ratios, and (b) varying ℓ.

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 4.2(a) 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 4.2(a) that for a fixed probability of success, the number of samples required increases with k/m, which agrees with the result in Theorem 4.2.1. In Figure 4.2(b), 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.

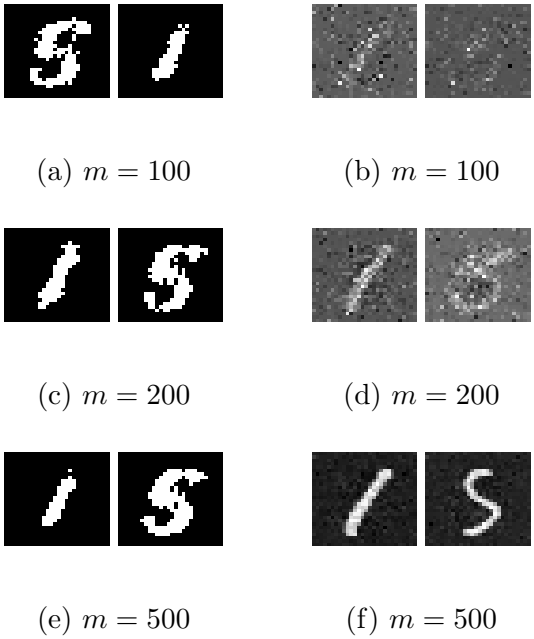
4.5.2 MNIST dataset

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 [42], 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 4.2.2. 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 4.3. The support size of each digit is roughly in the range $150 - 200$. It can be seen that Algorithm 2 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 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 4.3, where digits 1 and 5 have significant overlap.



(a) $m = 100$ (b) $m = 100$

(c) $m = 200$ (d) $m = 200$

(e) $m = 500$ (f) $m = 500$

Figure 4.3: Recovery performance of Algorithm 2 ((a),(c),(e)), and Group LASSO ((b),(d),(f)), with $n = 2000$ and varying m.

4.5.3 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 [86] 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.

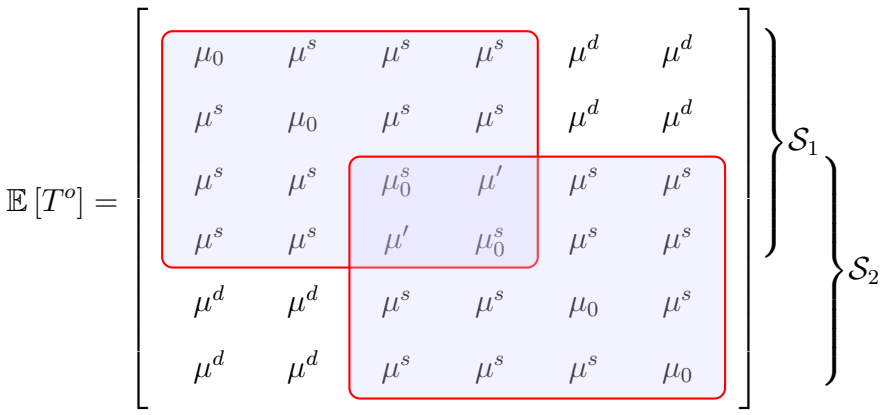


Figure 4.4: Block structure of the expected affinity matrix when the supports overlap, under appropriate permutation of rows and columns.

4.6 Overlapping supports

Our discussion till now focused on the case of disjoint supports. In this section, we describe an extension of our algorithm to handle intersecting supports when $\ell = 2$. In this setting, the expected affinity matrix has an overlapping block structure as shown in Figure 4.4, and the key step is to characterize the eigenvectors and eigenvalues of this matrix. In the $\ell = 2$ case, as we describe below, the sign pattern of the second leading eigenvector of T^o[3] determines the performance of the algorithm. In particular, the number of misclustered coordinates can be related to the eigenvalues of $\mathbb{E}[T^o]$ and an error term $\|T^o - \mathbb{E}[T^o]\|_{op}$. We characterize both these quantities and provide the performance guarantee in Theorem 4.6.1.

Let $k_{\text{un}} \stackrel{\text{def}}{=} |\mathcal{S}_1 \cup \mathcal{S}_2|$ and $k_{\text{int}} = |\mathcal{S}_1 \cap \mathcal{S}_2|$ denote the sizes of the union and intersection of the underlying supports, respectively. When $\mathcal{S}_1$ and $\mathcal{S}_2$ have a non empty intersection, the expected affinity matrix $\mathbb{E}[T^o]$ has a block structure under an unknown permutation of the rows and columns as depicted in Figure 4.4. It is well-known that the sign of the second leading eigenvector of $\mathbb{E}[T^o]$ can reveal the grouping of indices into the underlying blocks. In particular, the entries of the eigenvector at indices that belong exclusively to one of the supports will be strictly positive or strictly negative. For indices that lie

[3]To avoid confusion with the case of disjoint supports, we will denote the affinity matrix by T^o here.

Algorithm 3: Multiple support recovery: $\ell = 2$ case

Input: Measurements $\{Y_i\}_{i=1}^n$, Measurement matrices $\{\Phi_i\}_{i=1}^n$, k_{un}, k_{in}

Output: Support estimates $\hat{\mathcal{S}}_1$, $\hat{\mathcal{S}}_2$

1 If $\mathcal{S}_1$ and $\mathcal{S}_2$ disjoint, set $\tau = 0$, otherwise set $\tau = \frac{1}{2\sqrt{k}}$

2 Form variance estimates $a_1, \ldots, a_n$ with entries

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

for $j \in [n]$.

3 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

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

4 Construct affinity matrix $T \in \mathbb{R}^{k_{\text{un}} \times k_{\text{un}}}$ as

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

5 Compute (normalized) second leading eigenvector $\hat{v}_2$ of T. Declare

$$\hat{\mathcal{S}}_1 = \{i \in \hat{\mathcal{S}}_{\text{un}} : \hat{v}_{2,i} > -\tau\}$$
$$\hat{\mathcal{S}}_2 = \{i \in \hat{\mathcal{S}}_{\text{un}} : \hat{v}_{2,i} < \tau\}$$

in the intersection of the two supports, the entries will be zero. Since we will be using the eigenvector of the sample version T^o instead of $\mathbb{E}[T^o]$, we relax the requirement of entries being exactly zero and look for values in a small interval around zero, whereby our estimates for the two supports are

$$\hat{\mathcal{S}}_1 = \{i \in \hat{\mathcal{S}}_{\text{un}} : \hat{v}_{2,i} > -\tau\} \tag{4.26}$$

$$\hat{\mathcal{S}}_2 = \{i \in \hat{\mathcal{S}}_{\text{un}} : \hat{v}_{2,i} < \tau\}, \tag{4.27}$$

for an appropriate threshold $\tau > 0$. Thus coordinates for which $\hat{v}_{2,i} \in [-\tau, \tau]$ are included in both supports. The full algorithm is described in Algorithm 3, for which we have the following performance guarantee. For simplicity, we state our guarantee considering the total number of mislabeled coordinates as the recovery criterion and for fixed probability of error. As we saw before, it can be converted to a guarantee in terms of the sum metric for arbitrary error probability.

Theorem 4.6.1. *Let $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$ be the estimates in (4.26), with τ chosen as $c/\sqrt{k_{un}}$ for $c < 1$. Then, for every $\varepsilon > 0$,*

$$\Pr\left(\exists \sigma \in \mathcal{G}_2 \ s.t. \ |(\mathcal{S}_1 \Delta \hat{\mathcal{S}}_{\sigma(1)}) \cup (\mathcal{S}_2 \Delta \hat{\mathcal{S}}_{\sigma(2)})| \leq 2\varepsilon k_{un}\right) \geq \frac{2}{3} \tag{4.28}$$

provided $k_{int}/m \leq c' < 1$ and

$$n \geq \frac{C}{\varepsilon}\left(1 - \frac{k_{int}}{k_{un}}\right)\frac{k^4}{m^4}(\log k)^4 \log k_{un}. \tag{4.29}$$

Proof. Let $\nu_1^o \geq \cdots \geq \nu_{k_{un}}^o$ be the eigenvalues of $\mathbb{E}[T^o]$. Also, let $\hat{v}_2$ and v_2 denote the normalized second leading eigenvectors of T^o and $\mathbb{E}[T^o]$, respectively. From the Davis-Kahan theorem (stated in Section 4.8), we have

$$\sin(\angle(\hat{v}_2, v_2)) \leq \frac{2\|T^o - \mathbb{E}[T^o]\|_{op}}{\Delta^o_{\min}}, \tag{4.30}$$

where $\sin(\angle(x,y)) \stackrel{\text{def}}{=} \sqrt{1 - (x^\top y)^2/\|x\|_2^2\|y\|_2^2}$ and $\Delta^o_{\min} \stackrel{\text{def}}{=} \min\{\nu_1^o - \nu_2^o, \nu_2^o - \nu_3^o\}$. This

result can be quickly translated to a bound on $\|\hat{v}_2 - v_2\|_2$. Indeed, for any two unit vectors x and y such that $\sin(\angle(x,y)) \leq c$, we have $|x^\top y| \geq (x^\top y)^2 \geq 1 - c^2$. Using this, one can show that either $\|x-y\|_2^2 \leq 2c^2$ or $\|x+y\|_2^2 \leq 2c^2$ is guaranteed to hold. Thus, the Davis-Kahan theorem essentially states that

$$\min\{\|\hat{v}_2 - v_2\|_2, \|\hat{v}_2 + v_2\|_2\} \leq 2\sqrt{2}\frac{\|T^o - \mathbb{E}\left[T^o\right]\|_{op}}{\Delta^o_{\min}}, \tag{4.31}$$

that is, the true eigenvector and the sample eigenvector are close upto sign. Our goal will be to show that the quantity on the right is small, which would show that the error between the eigenvectors is also small. But first, we will relate the error between the eigenvectors to the error in the recovered supports.

We will identify events that lead to false alarm and missed detection errors. Towards that, we define

$$\mathcal{E}_1^{FA/MD} = \{i \in \mathcal{S}_{\text{un}} : v_{2,i} > 0, \hat{v}_{2,i} < -\tau\} \tag{4.32}$$

$$\mathcal{E}_2^{FA/MD} = \{i \in \mathcal{S}_{\text{un}} : v_{2,i} < 0, \hat{v}_{2,i} > \tau\}, \tag{4.33}$$

as events that lead to both false alarm and missed detection errors. In a similar way, we define the events

$$\mathcal{E}_3^{MD} = \{i \in \mathcal{S}_{\text{un}} : v_{2,i} = 0, \hat{v}_{2,i} \notin [-\tau, \tau]\} \tag{4.34}$$

$$\mathcal{E}_4^{FA} = \{i \in \mathcal{S}_{\text{un}} : v_{2,i} > 0, \hat{v}_{2,i} \in [-\tau, \tau]\} \tag{4.35}$$

$$\mathcal{E}_5^{FA} = \{i \in \mathcal{S}_{\text{un}} : v_{2,i} < 0, v_{2,i} \in [-\tau, \tau]\}. \tag{4.36}$$

Then, the error event is $\mathcal{E} = \mathcal{E}_1^{FA/MD} \cup \mathcal{E}_2^{FA/MD} \cup \mathcal{E}_3^{MD} \cup \mathcal{E}_4^{FA} \cup \mathcal{E}_5^{FA}$. Now, note that the entries of v_2 are either $1/\sqrt{k'}$, $-1/\sqrt{k'}$ or zero, where $k' = k_{\text{un}} - k_{\text{int}}$. The minimum on the left of (4.31) depends on the signs of entries of v_2 relative to $\mathcal{S}_1$ and $\mathcal{S}_2$. Note that from the description of the estimator, the entries of $\hat{v}_2$ are always positive on $\mathcal{S}_1 \backslash \mathcal{S}_2$ and negative on $\mathcal{S}_2 \backslash \mathcal{S}_1$. Assuming without loss of generality that $\mathcal{S}_1$ and $\mathcal{S}_2$ are such that the entries of v_2 are positive on $\mathcal{S}_1 \backslash \mathcal{S}_2$, negative on $\mathcal{S}_2 \backslash \mathcal{S}_1$, and zero on $\mathcal{S}_1 \cap \mathcal{S}_2$, we see that

the minimum in (4.31) is achieved by $\|\hat{v}_2 - v_2\|_2$, since the values add up on the error set $\mathcal{E}$ due to opposite signs. In particular, note that if $i \in \mathcal{E}_1^{FA/MD}$ or $i \in \mathcal{E}_2^{FA/MD}$, then $(v_{2,i} - \hat{v}_{2,i})^2 \geq 1/k'$, since the entries have opposite signs. On the other hand, if $i \in \mathcal{E}_3^{FA}$, then $(v_{2,i} - \hat{v}_i)^2 \geq \tau^2$. Finally, for $i \in \mathcal{E}_4^{FA}$ or $i \in \mathcal{E}_5^{FA}$, $(v_{2,i} - \hat{v}_i)^2 \geq (1/\sqrt{k'} - \tau)^2$. Choosing $\tau = c/\sqrt{k'}$ for $c < 1$, we get

$$\|v_2 - \hat{v}_2\|_2^2 \geq \sum_{i \in \mathcal{E}} (v_{2,i} - \hat{v}_{2,i})^2 \geq |\mathcal{E}| \frac{c'}{k'},$$

where $c' < 1$.

Similarly, when $\mathcal{S}_1$ and $\mathcal{S}_2$ are such that the entries of v_2 are negative on $\mathcal{S}_1 \backslash \mathcal{S}_2$, positive on $\mathcal{S}_2 \backslash \mathcal{S}_1$, and zero on $\mathcal{S}_1 \cap \mathcal{S}_2$, the minimum is achieved by $\|\hat{v}_2 + v_2\|_2^2 \geq c'|\mathcal{E}|/k'$. Combining these facts with (4.31), we see that if

$$\frac{\|T^o - \mathbb{E}[T^o]\|_{op}}{\Delta_{\min}} \leq \sqrt{\frac{k_{\text{un}}}{k'}} \varepsilon \tag{4.37}$$

with probability at least 2/3, then it implies that $|\mathcal{E}| \leq c\varepsilon k_{\text{un}}$ with probability at least 2/3. As before, we will use Theorem 4.4.11 to control $\|T^o - \mathbb{E}[T^o]\|_{op}$ by characterizing the spectrum of $\mathbb{E}[T^o]$. Using similar arguments as in the proof of Theorem 4.2.1, and letting $\varepsilon' = \sqrt{\varepsilon k_{\text{un}}/k'}$, we get

$$\Pr\left(\frac{\|T^o - \mathbb{E}[T^o]\|_{op}}{\Delta_{\min}^o} \geq \varepsilon' \right) \leq \frac{1}{3}, \tag{4.38}$$

provided

$$n \geq \frac{c}{\varepsilon'^2} \frac{\|\mathbb{E}[T^o]\|_{op}}{(\Delta_{\min}^o)^2} \left(\mathbb{E}\left[\|a_1\|_2^{\log k} \right] \right)^{\frac{2}{\log k}} \log k_{\text{un}}. \tag{4.39}$$

We now use the following lemma which characterizes the spectrum of $\mathbb{E}[T^o]$. The proof is provided in Section 4.8.7.

Lemma 4.6.2. *Under Assumptions 4.2.1 and 4.2.2, we have*

$$\|\mathbb{E}\left[T^o\right]\|_{op} \leq c\lambda_0^2 \frac{k^3}{m^2}.$$

Further, assuming $k_{int}/m \leq c'$ for some $c' < 1$, we have

$$\Delta^o_{\min} \geq c''\lambda_0^2 k.$$

Plugging the results from Lemma 4.6.2 into (4.39) and using Lemma 4.4.7 gives

$$n \geq \frac{c}{\varepsilon'^2}\frac{k^4}{m^4} k^{\frac{2}{\log k}} (\log k)^4 \log k_{\text{un}}. \tag{4.40}$$

Substituting $\varepsilon' = \sqrt{\varepsilon k_{un}/k'}$, and using similar arguments as in the proof of Theorem 4.2.1, we get

$$n \geq \frac{c}{\varepsilon}\left(1 - \frac{k_{\text{int}}}{k_{\text{un}}}\right)\frac{k^4}{m^4}(\log k)^4 \log k_{\text{un}}. \tag{4.41}$$

□

4.7 Discussion and Extensions

In our results in the initial part of this chapter, 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 $\ell = 2$, we showed that overlapping supports can be handled by a modification of the sign-based estimate. Given our current algorithm, a simple way to handle overlapping supports 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 our 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 Chapter 3 (with an additional $1/\ell$ factor), and obtaining a better lower bound seems difficult.

4.8 Remaining proofs

4.8.1 Proof of Lemma 4.4.4 (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

$$\min_{\sigma_t \in \mathcal{G}_\ell} \sum_{i=1}^{\ell} |\hat{\mathcal{S}}_{i,t} \Delta \hat{\mathcal{S}}_{\sigma_t(i),t'}| \le 2\varepsilon k \ell^2, \quad \forall\, t' \in \mathcal{I}. \tag{4.42}$$

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 $L \ge \lceil 8 \log \frac{1}{\delta} \rceil$ the estimator above constitutes an $(nL, 3\varepsilon, \delta)$-estimator

for $\Sigma_{k,\ell,d}$. Indeed, denoting

$$Z_t = \mathbb{1}\left(\exists\, \sigma \in \mathcal{G}_\ell \text{ s.t. } \sum_{i=1}^{\ell} |\mathcal{S}_i \Delta \hat{\mathcal{S}}_{\sigma(i),t}| \le \varepsilon k \ell^2\right),$$

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 $L \ge \lceil 8 \log \frac{1}{\delta} \rceil$, with probability exceeding $1-\delta$ there exist[4] $M \ge L/2+1$ indices $t_1,\ldots,t_M \in [L]$ and permutations $\sigma_1,\ldots,\sigma_M \in \mathcal{G}_\ell$ such that

$$\sum_{i=1}^{\ell} |\mathcal{S}_i \Delta \hat{\mathcal{S}}_{\sigma_j(i),t_j}| \le \varepsilon k \ell^2, \quad \forall\, j \in [M]. \tag{4.43}$$

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 (4.42) when (4.43) 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 (4.43) holds.

To that end, denote by $\mathcal{I}'$ the set $\{t_1,\ldots,t_M\}$ of indices satisfying (4.43), and recall the set $\mathcal{I}$ found by our estimation procedure earlier. Then, when $|\mathcal{I}'| \ge L/2+1$, which holds with probability exceeding $1-\delta$,

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

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

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

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

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

$$\sum_{i=1}^{\ell} |\mathcal{S}_i \Delta \overline{\mathcal{S}}_{\sigma'(i)}| \leq 3\varepsilon k \ell^2,$$

which completes the proof. □

4.8.2 Proof of Lemma 4.4.7

As noted in the proof of Theorem 4.2.1, 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 $q \geq 2$,

$$\begin{aligned}
\mathbb{E}\left[\max_{i\in[n]} \|a_i\|_2^2\right] &= \mathbb{E}\left[\left(\max_{i\in[n]} \|a_i\|_2^q\right)^{\frac{2}{q}}\right] \\
&\leq \mathbb{E}\left[\left(\sum_{i=1}^{n} \|a_i\|_2^q\right)^{\frac{2}{q}}\right] \\
&\leq \left(\mathbb{E}\left[\sum_{i=1}^{n} \|a_i\|_2^q\right]\right)^{\frac{2}{q}} \\
&= n^{\frac{2}{q}} \left(\mathbb{E}\left[\|a_1\|_2^q\right]\right)^{\frac{2}{q}},
\end{aligned}$$

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

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

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

$$\begin{aligned}\|a\|_2^{2q} = \Big(\sum_{i\in\mathcal{S}_{\text{un}}} a_i^2\Big)^q &= \Big(\sum_{i\in\mathcal{S}_{\text{un}}} (\Phi_i^\top \Phi_{\mathcal{S}} X_{\mathcal{S}})^4\Big)^q \\ &= \Big(\sum_{i\in\mathcal{S}_{\text{un}}} (\alpha_i^\top X_{\mathcal{S}})^4\Big)^q \\ &= \Big(\sum_{i\in\mathcal{S}_{\text{un}}} (X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\Big)^q,\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{i\in\mathcal{S}_{\text{un}}} (X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\Big)^q \Big| \Phi\Big] &= (k\ell)^q \mathbb{E}\Big[\Big(\frac{1}{k\ell}\sum_{i\in\mathcal{S}_{\text{un}}} (X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\Big)^q \Big| \Phi\Big] \\ &\leq (k\ell)^{q-1} \sum_{i\in\mathcal{S}_{\text{un}}} \mathbb{E}\left[(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^{2q} | \Phi\right], \qquad (4.44)\end{aligned}$$

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 Section 4.9 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.* [68]) to get

$$\mathbb{P}(|X_{\mathcal{S}}^\top A_i X_{\mathcal{S}} - \mu| \geq t | \Phi) \leq 2\exp\Big(-\min\Big\{\frac{t^2}{\lambda_0^2 \|A_i\|_F^2}, \frac{t}{\lambda_0 \|A_i\|_{op}}\Big\}\Big),$$

where $\mu = \mathbb{E}\left[X_{\mathcal{S}}^\top A_i X_{\mathcal{S}} | \Phi\right] = \lambda_0 \|\alpha_i\|_2^2$. Lemma 4.9.1 in Section 4.9 can now be used to bound the moment in (4.44). Specifically, we get

$$\begin{aligned}\mathbb{E}[(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^{2q} | \Phi] &\leq 2q \cdot (16)^q \Big(\Gamma(q)\lambda_0^{2q}\|A_i\|_F^{2q} + \Gamma(2q)\lambda_0^{2q}\|A_i\|_{op}^{2q}\Big) + 2^{2q}\mu^{2q} \\ &\leq 3q \cdot (16)^q \Gamma(2q) \lambda_0^{2q} \|\alpha_i\|_2^{4q},\end{aligned}$$

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

$$\mathbb{E}\left[(X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^{2q}\right] \le c_q' \Gamma(2q) \lambda_0^{2q} \mathbb{E}\left[\|\alpha_i\|_2^{4q}\right], \tag{4.45}$$

where $c_q' = 3q \cdot (16)^q$. Thus, combining the result above with (4.44), we get

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i\in\mathcal{S}_{\text{un}}} (X_{\mathcal{S}}^\top A_i X_{\mathcal{S}})^2\right)^q\right] &\le c_q' \Gamma(2q) \lambda_0^{2q} (k\ell)^q \sum_{i\in\mathcal{S}_{\text{un}}} \mathbb{E}\left[\|\alpha_i\|_2^{4q}\right] \\ &= c_q' \Gamma(2q) \lambda_0^{2q} (k\ell)^q \left(\sum_{i\in\mathcal{S}} \mathbb{E}\left[\|\alpha_i\|_2^{4q}\right] + \sum_{i\in\mathcal{S}_{\text{un}}\setminus\mathcal{S}} \mathbb{E}\left[\|\alpha_i\|_2^{4q}\right]\right).\end{aligned} \tag{4.46}$$

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

$$\begin{aligned}\mathbb{E}[\|\alpha_i\|_2^{4q}] &= \mathbb{E}\left[\left(\|\Phi_i\|_2^4 + \sum_{j\in\mathcal{S}\setminus\{i\}} (\Phi_i^\top \Phi_j)^2\right)^{2q}\right] \\ &\le 2^{2q}\left(\mathbb{E}\left[\|\Phi_i\|_2^{8q}\right] + \mathbb{E}\left[\left(\sum_{j\in\mathcal{S}\setminus\{i\}} (\Phi_i^\top \Phi_j)^2\right)^{2q}\right]\right),\end{aligned}$$

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

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

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 $c'' = 8$ [60, Lemma D.2]. This gives, using Lemma 4.9.1,

$$\begin{aligned}\mathbb{E}\left[(\|\Phi_i\|_2^2)^{4q}\right] &\le 2q(16)^q \left(\Gamma(2q)\frac{c'^{2q}}{m^{2q}} + \Gamma(4q)\frac{c''^{4q}}{m^{4q}}\right) + \left(\mathbb{E}\left[\|\Phi_i\|_2^2\right]\right)^{4q} \\ &\le 4q(16)^q c'^{2q} \Gamma(4q)\frac{1}{m^{2q}} + 1,\end{aligned}$$

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 4.9.1 again, we get

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

Combining these results and substituting into (4.46), we get

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i\in\mathcal{S}_{\text{un}}}(X_\mathcal{S}^\top A_i X_\mathcal{S})^2\right)^q\right] &\leq c'_q\Gamma(2q)\lambda_0^{2q}(k\ell)^{q-1}\left(\sum_{i\in S}\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right] + \sum_{i\in\mathcal{S}_{\text{un}}\setminus\mathcal{S}}\mathbb{E}\left[\|\alpha_i\|_2^{4q}\right]\right) \\
&\leq 5c'^2_q c'^{2q}\Gamma(2q)\lambda_0^{2q}(k\ell)^{q-1}\left(k\Gamma(2q)\left(\frac{k}{m}\right)^{2q} + (k\ell-k)\Gamma(2q)\left(\frac{k}{m}\right)^{2q}\right) \\
&= 5c'^2_q c'^{2q}(\Gamma(2q))^2\lambda_0^{2q}\left(\frac{k\sqrt{k\ell}}{m}\right)^{2q}.
\end{aligned}$$

Rescaling the exponent, we get

$$\begin{aligned}
\mathbb{E}\left[\|a\|_2^q\right] &= \mathbb{E}\left[\left(\sum_{i\in\mathcal{S}_{\text{un}}}(X_\mathcal{S}^\top A_i X_\mathcal{S})^2\right)^{\frac{q}{2}}\right] \\
&\leq 5c_{q/2}^2 c'^q(\Gamma(q))^2\lambda_0^q\left(\frac{k\sqrt{k\ell}}{m}\right)^q
\end{aligned}$$

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

4.8.3 Proof of Lemma 4.4.9

(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

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

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

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

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 [65, 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

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

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,

$$WBW^\top = WH\Lambda(WH)^\top,$$

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

$$WBW^\top \cdot WH = WH \cdot \Lambda,$$

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 $w^i H = w^j H$. 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 $w^i H = w^j H$, which implies $v^i = v^j$.

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

$$\begin{aligned} \|v^i - v^j\|_2 &= \|(w^i - w^j)H\|_2 \\ &\geq \sqrt{2}\nu_{\min}(H), \end{aligned}$$

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,

$$\begin{aligned} \min_{\|x\|_2=1} \|x^\top H\|_2^2 &= \min_{\|x\|_2=1} x^\top HH^\top x \\ &= \min_{\|x\|_2=1} x^\top (WW^\top)^{-1} x \\ &= \frac{1}{k}, \end{aligned}$$

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

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

4.8.4 Proof of Theorem 4.4.11

The proof is similar to that of [66], and we highlight the steps needed to extend the result to general A. In particular, following similar arguments as in [66], it can be shown that

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top - A\right\|_{op}\right] \leq c\frac{\sqrt{\log N}}{n}\sqrt{\mathbb{E}\left[\max_{i\in[n]}\|Z_i\|_2^2\right]}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{n} Z_i Z_i^\top\right\|_{op}\right]}, \tag{4.47}$$

Now,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{n} Z_i Z_i^\top\right\|_{op}\right] &\leq n\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top - A\right\|_{op} + \|A\|_{op}\right] \\ &= n(\beta + \|A\|_{op}),\end{aligned} \tag{4.48}$$

where $\beta \stackrel{\text{def}}{=} \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top - A\right\|_{op}\right]$. It follows from (4.47) and (4.48) that

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

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

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

which completes the proof.

4.8.5 Proof of Lemma 4.4.3

Our goal is to compute the expected value of the affinity matrix, denoted $\mathbb{E}[T]$, and we will do so by first conditioning on the measurement ensemble Φ_1^n and noting that each entry of T is then of the form $(X^\top A X)^2$, where X is subgaussian and A is a fixed matrix (given Φ_1^n). This conditional expectation can be calculated using Lemma 4.9.2. The next step is to average over the distribution of Φ_1^n, and our analysis will require the moment

assumptions on the entries of Φ_1^n described in Assumption 2. Although each entry of $\mathbb{E}[T]$ can be explicitly characterized in terms of the system parameters, we will sometimes only mention the leading terms. In fact, the analysis of our algorithm in Theorem 1 only requires an upper bound on the diagonal entries and tight upper and lower bounds on the off diagonal entries of $\mathbb{E}[T]$.

Specifically, by the definition of T from (4.49), we note that

$$\mathbb{E}[T_{uv}] = \frac{1}{n}\sum_{j=1}^{n}(\Phi_{ju}^{\top}\Phi_j X_j)^2 \cdot (\Phi_{jv}^{\top}\Phi_j X_j)^2, \tag{4.49}$$

for $(u,v) \in \mathcal{S}_{\text{un}} \times \mathcal{S}_{\text{un}}$. The expectation in the expression above is over the joint distribution of X_1^n, Φ_1^n and the labels G_1^n (generating samples from the mixture $\mathrm{P}_{\mathcal{S}} = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathrm{P}^{(i)}$ described in Section II in the main file can be thought of as drawing the label G uniformly from $[\ell]$, and conditioned on $G=g$, drawing a sample from $\mathrm{P}^{(g)}$). We will first condition on the labels (or, equivalently, on the random subsets $\{I_1,\ldots,I_\ell\}$ defined as $I_i \stackrel{\text{def}}{=} \{j \in [n] : \texttt{supp}(X_j) = \mathcal{S}_i\}$ and on the measurement matrices. We focus on a single summand in (4.49), and drop the dependence on the sample index j. With a slight abuse of notation, we let $\mathcal{S} = \text{supp}(X)$ denote the support of the sample we focus on and note that

$$\mathbb{E}_X\left[(\Phi_u^{\top}\Phi X)^2 \cdot (\Phi_v^{\top}\Phi X)^2 | \Phi, G\right] = \mathbb{E}_X\left[(X_{\mathcal{S}}^{\top}\alpha_u\alpha_v^{\top}X_{\mathcal{S}})^2 | \Phi, G\right],$$

where, $\alpha_u \stackrel{\text{def}}{=} \Phi_{\mathcal{S}}^{\top}\Phi_u$, $u \in \mathcal{S}_{\text{un}}$. We can now use Lemma 4.9.2 to get

$$\mathbb{E}_X\left[(X_{\mathcal{S}}^{\top}\alpha_u\alpha_v^{\top}X_{\mathcal{S}})^2 | \Phi, G\right] = \rho\sum_{i\in\mathcal{S}}\alpha_{ui}^2\alpha_{vi}^2 + \lambda_0^2\sum_{i\neq j}\alpha_{ui}^2\alpha_{vj}^2 + \lambda_0^2\sum_{i\neq j}\alpha_{ui}\alpha_{vi}\alpha_{uj}\alpha_{vj}, \tag{4.50}$$

where recall $\lambda_0 = \mathbb{E}[X_i^2]$ and $\rho = \mathbb{E}[X_i^4]$. We will first handle the $u=v$ case, which will be used to compute the diagonal entries of the mean matrix. We have, for every $u \in \mathcal{S}_{\text{un}}$,

$$\mathbb{E}_{X,\Phi}\left[(X_{\mathcal{S}}^{\top}\alpha_u\alpha_u^{\top}X_{\mathcal{S}})^2 | G\right] = \rho\mathbb{E}_{\Phi}\left[\sum_{i\in\mathcal{S}}\alpha_{ui}^4 | G\right] + 2\lambda_0^2\mathbb{E}_{\Phi}\left[\sum_{i\neq j}\alpha_{ui}^2\alpha_{uj}^2 | G\right]$$

$$= \rho\mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}}(\Phi_u^\top\Phi_i)^4|G\right] + 2\lambda_0^2\mathbb{E}_\Phi\left[\sum_{i\neq j}(\Phi_u^\top\Phi_i)^2(\Phi_u^\top\Phi_j)^2|G\right].$$

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

$$\begin{aligned}
\mu_0^s &\stackrel{\text{def}}{=} \mathbb{E}_{X,\Phi}\left[(X_\mathcal{S}^\top\alpha_u\alpha_u^\top X_\mathcal{S})^2|G\right] \\
&= \rho\mathbb{E}_\Phi\left[\|\Phi_u\|_2^8 + \sum_{i\in\mathcal{S}\setminus\{u\}}(\Phi_u^\top\Phi_i)^4|G\right] + 2\lambda_0^2\mathbb{E}_\Phi\left[2\|\Phi_u\|_2^4\sum_{i\in\mathcal{S}\setminus\{u\}}(\Phi_u^\top\Phi_i)^2 + \sum_{i\neq j}(\Phi_u^\top\Phi_i)^2(\Phi_u^\top\Phi_j)^2|G\right] \\
&\leq c\rho\left(1+\frac{k-1}{m^2}\right) + c'\lambda_0^2\left(\frac{k-1}{m} + \frac{(k-1)(k-2)}{m^2}\right), \qquad (4.51)
\end{aligned}$$

where we used Lemma 4.9.2 in the second step and Lemma 4.9.4 in the third step, and retained the leading terms.

When $u \in \mathcal{S}_{\text{un}}\setminus\mathcal{S}$, using Lemmas 4.9.2 and 4.9.4 once again, we have

$$\begin{aligned}
\mu_0^d &\stackrel{\text{def}}{=} \mathbb{E}_{X,\Phi}\left[(X_\mathcal{S}^\top\alpha_u\alpha_u^\top X_\mathcal{S})^2|G\right] \\
&= \rho\mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}}(\Phi_u^\top\Phi_i)^4|G\right] + 2\lambda_0^2\mathbb{E}_\Phi\left[\sum_{i\neq j}(\Phi_u^\top\Phi_i)^2(\Phi_u^\top\Phi_j)^2|G\right] \\
&\leq c\rho\left(\frac{k}{m^2}\right) + c'\lambda_0^2\frac{k(k-1)}{m^2}. \qquad (4.52)
\end{aligned}$$

We now use these results to bound the diagonal entries of the mean matrix $\mathbb{E}[T]$. Using (4.49), (4.51) and (4.52), we see that for $u \in \mathcal{S}_1$,

$$\begin{aligned}
\mu_0 \stackrel{\text{def}}{=} \mathbb{E}[T_{uu}] &= \mathbb{E}_G\left[\mathbb{E}_{X,\Phi}\left[\frac{1}{n}\left(\sum_{j\in I_1}(\Phi_{ju}^\top\Phi_j X_j)^4 + \cdots + \sum_{j\in I_\ell}(\Phi_{ju}^\top\Phi_j X_j)^4\right)\middle| G\right]\right] \\
&= \mathbb{E}_G\left[\frac{1}{n}\left(|I_1|\mu_0^s + \sum_{i=2}^{\ell}|I_i|\mu_0^d\right)\right] \\
&= \frac{1}{\ell}\mu_0^s + \frac{\ell-1}{\ell}\mu_0^d \\
&\leq \frac{c}{\ell}\left\{\rho\left(1+\frac{k-1}{m^2}\right) + \lambda_0^2\left(\frac{k-1}{m} + \frac{(k-1)(k-2)}{m^2}\right)\right\} \\
&\quad + \frac{c(\ell-1)}{\ell}\left\{\rho\left(\frac{k}{m^2}\right) + \lambda_0^2\frac{k(k-1)}{m^2}\right\}, \qquad (4.53)
\end{aligned}$$

where we used $\mathbb{E}_G[|I_i|] = n/\ell$ for all $i \in [\ell]$, under the uniform mixture assumption. The same result holds for $u \in \mathcal{S}_i$ for any $i \in [\ell]$.

The next step is to bound the off diagonal entries of $\mathbb{E}[T]$. Continuing from (4.50), we will handle each of the three terms separately. For each of these terms, we will consider the case when both u and v belong to the same support, and when they belong to different supports. Overall, these calculations highlight the block structure of $\mathbb{E}[T]$, with the diagonal entries all being equal, and the off diagonal entries taking two different values based on whether the indices belong to the same support or not.

For the first term in (4.50), when $(u, v) \in \mathcal{S} \times \mathcal{S}$, $u \neq v$, we have

$$\begin{aligned}\mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}} \alpha_{ui}^2\alpha_{vi}^2|G\right] =& \mathbb{E}_\Phi\left[\|\Phi_u\|_2^4(\Phi_u^\top\Phi_v)^2|G\right] + \mathbb{E}_\Phi\left[\|\Phi_v\|_2^4(\Phi_u^\top\Phi_v)^2|G\right] \\ &+ \mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}\setminus\{u\}\cup\{v\}} (\Phi_i^\top\Phi_u)^2(\Phi_i^\top\Phi_v)^2\middle|G\right] \\ =& \frac{2}{m}\left(1 + \frac{3}{m}(c_2-1) + \frac{1}{m^2}(c_3 - 3c_2 + 2)\right) + \frac{k-2}{m^2}\left(1 + \frac{1}{m}(c_2-1)\right) \stackrel{\text{def}}{=} \gamma_1^s,\end{aligned} \tag{4.54}$$

using Lemma 4.9.4. On the other hand, when $(u, v) \in \mathcal{S} \times \mathcal{S}_{\text{un}}\setminus\mathcal{S}$, we have

$$\begin{aligned}\mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}} \alpha_{ui}^2\alpha_{vi}^2|G\right] =& \mathbb{E}_\Phi\left[\|\Phi_u\|_2^4(\Phi_u^\top\Phi_v)^2|G\right] + \mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}\setminus\{u\}} (\Phi_i^\top\Phi_u)^2(\Phi_i^\top\Phi_v)^2\middle|G\right] \\ =& \frac{1}{m}\left(1 + \frac{3}{m}(c_2-1) + \frac{1}{m^2}(c_3 - 3c_2 + 2)\right) + \frac{k-1}{m^2}\left(1 + \frac{1}{m}(c_2-1)\right) \\ \stackrel{\text{def}}{=}& \gamma_1^{sd}.\end{aligned} \tag{4.55}$$

The same result holds when $(u, v) \in \mathcal{S}_{\text{un}}\setminus\mathcal{S} \times \mathcal{S}$. Finally, when $(u, v) \in \mathcal{S}_{\text{un}}\setminus\mathcal{S} \times \mathcal{S}_{\text{un}}\setminus\mathcal{S}$,

$$\begin{aligned}\mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}} \alpha_{ui}^2\alpha_{vi}^2|G\right] &= \mathbb{E}_\Phi\left[\sum_{i\in\mathcal{S}} (\Phi_i^\top\Phi_u)^2(\Phi_i^\top\Phi_v)^2|G\right] \\ &= \frac{k}{m^2}\left(1 + \frac{1}{m}(c_2-1)\right) \stackrel{\text{def}}{=} \gamma_1^d.\end{aligned} \tag{4.56}$$

For the second term in (4.50), when $(u,v) \in \mathcal{S} \times \mathcal{S}$,

$$\begin{aligned}
\mathbb{E}_\Phi\left[\sum_{i\neq j}\alpha_{ui}^2\alpha_{vj}^2|G\right] &= \mathbb{E}_\Phi\left[\|\Phi_u\|_2^4\|\Phi_v\|_2^4 + (\Phi_u^\top\Phi_v)^4 + \|\Phi_u\|_2^4\sum_{i\in\mathcal{S}\setminus\{u\}\cup\{v\}}(\Phi_v^\top\Phi_i)^2\middle|G\right] \\
&\quad + \mathbb{E}_\Phi\left[\|\Phi_v\|_2^4\sum_{i\in\mathcal{S}\setminus\{u\}\cup\{v\}}(\Phi_u^\top\Phi_i)^2 + (\Phi_u^\top\Phi_v)^2\sum_{i\in\mathcal{S}\setminus\{u\}\cup\{v\}}(\Phi_v^\top\Phi_i)^2\middle|G\right] \\
&\quad + \mathbb{E}_\Phi\left[(\Phi_u^\top\Phi_v)^2\sum_{i\in\mathcal{S}\setminus\{u\}\cup\{v\}}(\Phi_u^\top\Phi_i)^2 + \sum_{\substack{i,j\in\mathcal{S}\setminus\{u\}\cup\{v\}\\ i\neq j}}(\Phi_u^\top\Phi_i)^2\cdot(\Phi_v^\top\Phi_j)^2\middle|G\right] \\
&= \left(1+\frac{1}{m}(c_2-1)\right)^2 + \left(\frac{2}{m^2}+\frac{1}{m^3}(c_2^2-2)\right) + 2\left(1+\frac{1}{m}(c_2-1)\right)\frac{k-2}{m} \\
&\quad + 2\frac{(k-2)}{m^2}\left(1+\frac{1}{m}(c_2-1)\right) + \frac{(k-2)(k-3)}{m^2} \stackrel{\text{def}}{=} \gamma_2^s, \qquad (4.57)
\end{aligned}$$

where we used Lemma 4.9.4 in the second step. When $(u,v) \in \mathcal{S} \times \mathcal{S}_{\text{un}}\backslash\mathcal{S}$,

$$\begin{aligned}
\mathbb{E}_\Phi\left[\sum_{i\neq j}\alpha_{ui}^2\alpha_{vj}^2|G\right] &= \mathbb{E}_\Phi\left[\|\Phi_u\|_2^4\sum_{i\in\mathcal{S}\setminus\{u\}}(\Phi_v^\top\Phi_i)^2|G\right] + \mathbb{E}_\Phi\left[(\Phi_u^\top\Phi_v)^2\sum_{i\in\mathcal{S}\setminus\{u\}}(\Phi_u^\top\Phi_i)^2|G\right] \\
&\quad + \mathbb{E}_\Phi\left[\sum_{\substack{i,j\in\mathcal{S}\setminus\{u\}\\ j\neq i}}(\Phi_u^\top\Phi_i)^2\cdot(\Phi_v^\top\Phi_j)^2|G\right] \\
&= \left(1+\frac{1}{m}(c_2-1)\right)\frac{k-1}{m} + \frac{(k-1)}{m^2}\left(1+\frac{1}{m}(c_2-1)\right) + \frac{(k-1)(k-2)}{m^2} \\
&\stackrel{\text{def}}{=} \gamma_2^{sd}, \qquad (4.58)
\end{aligned}$$

and the same expression holds when $(u,v) \in \mathcal{S}_{\text{un}}\backslash\mathcal{S} \times \mathcal{S}$. When $(u,v) \in \mathcal{S}_{\text{un}}\backslash\mathcal{S} \times \mathcal{S}_{\text{un}}\backslash\mathcal{S}$,

$$\mathbb{E}_\Phi\left[\sum_{i\neq j}\alpha_{ui}^2\alpha_{vj}^2|G\right] = \mathbb{E}_\Phi\left[\sum_{\substack{i,j\in\mathcal{S}\\ j\neq i}}(\Phi_u^\top\Phi_i)^2\cdot(\Phi_v^\top\Phi_j)^2|G\right] = \frac{k(k-1)}{m^2} \stackrel{\text{def}}{=} \gamma_2^d, \qquad (4.59)$$

Finally, for the third term in (4.50), when $(u,v) \in \mathcal{S} \times \mathcal{S}$,

$$\begin{aligned}
\mathbb{E}_\Phi\left[\sum_{i\neq j}\alpha_{ui}\alpha_{vi}\alpha_{uj}\alpha_{vj}|G\right] =& \mathbb{E}_\Phi\left[\|\Phi_u\|_2^2\Phi_u^\top\Phi_v\cdot\|\Phi_v\|_2^2\Phi_u^\top\Phi_v|G\right] \\
&+ \mathbb{E}_\Phi\left[\|\Phi_u\|_2^2\Phi_u^\top\Phi_v\sum_{j\in\mathcal{S}\backslash\{u\}\cup\{v\}}(\Phi_u^\top\Phi_j)\cdot(\Phi_v^\top\Phi_j)|G\right] \\
&+ \mathbb{E}_\Phi\left[\|\Phi_v\|_2^2\Phi_u^\top\Phi_v\sum_{j\in\mathcal{S}\backslash\{u\}\cup\{v\}}(\Phi_u^\top\Phi_j)\cdot(\Phi_v^\top\Phi_j)|G\right] \\
&+ \mathbb{E}_\Phi\left[\sum_{\substack{i,j\in\mathcal{S}\backslash\{u\}\cup\{v\}\\ j\neq i}}(\Phi_u^\top\Phi_i)(\Phi_v^\top\Phi_i)(\Phi_u^\top\Phi_j)(\Phi_v^\top\Phi_j)|G\right] \\
=& \frac{1}{m}\left(1+\frac{c_2-1}{m}\right)^2 + \frac{2(k-2)}{m^2}\left(1+\frac{c_2-1}{m}\right) + \frac{(k-2)(k-3)}{m^3} \\
\stackrel{\text{def}}{=}& \gamma_3^s. \qquad (4.60)
\end{aligned}$$

When $(u,v) \in \mathcal{S} \times \mathcal{S}_{\text{un}}\backslash\mathcal{S}$,

$$\begin{aligned}
\mathbb{E}_\Phi\left[\sum_{i\neq j}\alpha_{ui}\alpha_{vi}\alpha_{uj}\alpha_{vj}|G\right] =& \mathbb{E}_\Phi\left[\|\Phi_u\|_2^2\Phi_u^\top\Phi_v\sum_{j\in\mathcal{S}\backslash\{u\}}(\Phi_u^\top\Phi_j)\cdot(\Phi_v^\top\Phi_j)|G\right] \\
&+ \mathbb{E}_\Phi\left[\sum_{\substack{i,j\in\mathcal{S}\backslash\{u\}\\ j\neq i}}(\Phi_u^\top\Phi_i)(\Phi_u^\top\Phi_j)(\Phi_v^\top\Phi_i)(\Phi_v^\top\Phi_j)\Big|G\right] \\
=& \frac{(k-1)}{m^2}\left(1+\frac{c_2-1}{m}\right) + \frac{(k-1)(k-2)}{m^3} \stackrel{\text{def}}{=} \gamma_3^{sd}, \qquad (4.61)
\end{aligned}$$

and the same expression holds when $(u,v) \in \mathcal{S}_{\text{un}}\backslash\mathcal{S} \times \mathcal{S}$. When $(u,v) \in \mathcal{S}_{\text{un}}\backslash\mathcal{S} \times \mathcal{S}_{\text{un}}\backslash\mathcal{S}$,

$$\begin{aligned}
\mathbb{E}_\Phi\left[\sum_{i\neq j}\alpha_{ui}\alpha_{vi}\alpha_{uj}\alpha_{vj}|G\right] &= \mathbb{E}_\Phi\left[\sum_{\substack{i,j\in\mathcal{S}\\ j\neq i}}(\Phi_u^\top\Phi_i)(\Phi_u^\top\Phi_j)(\Phi_v^\top\Phi_i)(\Phi_v^\top\Phi_j)\Big|G\right] \\
&= \frac{k(k-1)}{m^3} \stackrel{\text{def}}{=} \gamma_3^d, \qquad (4.62)
\end{aligned}$$

We have thus computed the expected values of each of the three terms in (4.50).

Thus, combining (4.54), (4.57) and (4.60) and using (4.50) and (4.49), we have for $(u,v) \in \mathcal{S}_1 \times \mathcal{S}_1$, $u \neq v$,

$$\begin{aligned}\mathbb{E}\left[T_{uv}\right] =& \mathbb{E}_G\Bigg[\frac{1}{n}\Bigg(\sum_{j\in I_1}\rho\gamma_1^s + \lambda_0^2(\gamma_2^s+\gamma_3^s) + \sum_{j\in I_2}\rho\gamma_1^d + \lambda_0^2(\gamma_2^d+\gamma_3^d) \\ &+ \cdots + \sum_{j\in I_\ell}\rho\gamma_1^d + \lambda_0^2(\gamma_2^d+\gamma_3^d)\Bigg)\Bigg] \\ =& \frac{1}{\ell}\Bigg(\rho\gamma_1^s + \lambda_0^2(\gamma_2^s+\gamma_3^s)\Bigg) + \frac{\ell-1}{\ell}\Bigg(\rho\gamma_1^d + \lambda_0^2(\gamma_2^d+\gamma_3^d)\Bigg) \stackrel{\text{def}}{=} \mu_s, \end{aligned} \tag{4.63}$$

where again we used $\mathbb{E}_G\left[|I_i|\right] = n/\ell$ for all $i \in [\ell]$. This holds for $(u,v) \in \mathcal{S}_i \times \mathcal{S}_i$, for every $i \in [\ell]$.

For the case when $(u,v) \in \mathcal{S}_1 \times \mathcal{S}_2$ or when $(u,v) \in \mathcal{S}_2 \times \mathcal{S}_1$,

$$\begin{aligned}\mathbb{E}\left[T_{uv}\right] =& \mathbb{E}_G\Bigg[\frac{1}{n}\Bigg(\sum_{j\in I_1}\rho\gamma_1^{sd} + \lambda_0^2(\gamma_2^{sd}+\gamma_3^{sd}) + \sum_{j\in I_2}\rho\gamma_1^{sd} + \lambda_0^2(\gamma_2^{sd}+\gamma_3^{sd}) \\ &+ \sum_{j\in I_3}\rho\gamma_1^d + \lambda_0^2(\gamma_2^d+\gamma_3^d) + \cdots + \sum_{j\in I_\ell}\rho\gamma_1^d + \lambda_0^2(\gamma_2^d+\gamma_3^d)\Bigg)\Bigg] \qquad (4.64) \\ =& \frac{2}{\ell}\Bigg(\rho\gamma_1^{sd} + \lambda_0^2(\gamma_2^{sd}+\gamma_3^{sd})\Bigg) + \frac{\ell-2}{\ell}\Bigg(\rho\gamma_1^d + \lambda_0^2(\gamma_2^d+\gamma_3^d)\Bigg) \stackrel{\text{def}}{=} \mu_d. \qquad (4.65)\end{aligned}$$

Again, the same expression holds for $\mathbb{E}\left[T_{uv}\right]$ whenever $(u,v) \in \mathcal{S}_i \times \mathcal{S}_j$, $i,j \in [\ell]$, $i \neq j$. The mean matrix $\mathbb{E}\left[T\right]$ thus has a block structure with μ_0 on the diagonal, μ_s on the remaining entries in the diagonal blocks and μ_d on the off diagonal blocks as depicted in Figure 4.1.

4.8.6 Proof of Lemma 4.4.6

Using the structure of $\mathbb{E}\left[T\right]$ derived in Lemma 4.4.3, we have,

$$\begin{aligned}\|\mathbb{E}\left[T\right]\|_{op} &= \mu_0 + (k-1)\mu_s + k(\ell-1)\mu_d \\ &\leq \rho\frac{k^2\ell}{m^2} + \lambda_0^2\frac{k^3\ell}{m^2},\end{aligned}$$

where we have used the definitions in (4.53), (4.63) and (4.65), and simplified.

For the eigengap computation, we first note from the definitions in (4.63) and (4.65) that

$$\begin{aligned}\mu_s - \mu_d =& \frac{\rho}{\ell}(\gamma_1^s + \gamma_1^d - 2\gamma_1^{sd}) + \frac{\lambda_0^2}{\ell}(\gamma_2^s + \gamma_2^d - 2\gamma_2^{sd} + \gamma_3^s + \gamma_3^d - 2\gamma_3^{sd}) \\ =& \frac{\rho}{\ell} \cdot 0 + \frac{\lambda_0^2}{\ell}\left\{\left(1 + \frac{c_2 - 1}{m}\right)^2 + \frac{1}{m^2}\left(2 + \frac{c_2^2 - 2}{m}\right)\right. \\ & \left. + \frac{1}{m}\left(1 + \frac{c_2 - 1}{m}\right)^2 - \frac{2}{m}\left(1 + \frac{c_2 - 1}{m}\right)\left(1 + \frac{2}{m}\right) + \frac{4}{m^2}\right\} \\ \geq& \frac{\lambda_0^2}{\ell}.\end{aligned}$$

We therefore have,

$$\Delta_\ell = \nu_\ell - \nu_{\ell+1} = k(\mu_s - \mu_d) \geq \frac{\lambda_0^2 k}{\ell}.$$

4.8.7 Proof of Lemma 4.6.2

In this proof, we characterize the spectrum of the expected affinity matrix for $\ell = 2$ when $\mathcal{S}_1$ and $\mathcal{S}_2$ have a non empty intersection. We will express the expected clustering ,atrix $\mathbb{E}[T^o]$ as the sum of $\mathbb{E}[T]$ in the disjoint case and a "small" perturbation. This will allow us to use results from the case with disjoint supports.

Using the same calculation as in Lemma 4.4.3, we can show that if $u \in \mathcal{S}_1 \backslash \mathcal{S}_2$ or $u \in \mathcal{S}_2 \backslash \mathcal{S}_1$, then

$$\mathbb{E}[T^o_{uu}] = \frac{1}{2}(\mu_0^s + \mu_0^d) = \mu_0, \tag{4.66}$$

whereas if $u \in \mathcal{S}_1 \cap \mathcal{S}_2$, then

$$\mathbb{E}[T^o_{uu}] = \mu_0^s = 2\mu_0 - \mu_0^d. \tag{4.67}$$

Thus, the diagonal entries are not all equal in this case, and have larger value along coordinates in the intersection (since $\mu_0^s \geq \mu_0^d$). For the off-diagonal entries, if $(u, v) \in$

$$M_{\text{err}} = \begin{bmatrix} 0 & 0 & 0 & \beta_2 & 0 & 0 \\ 0 & 0 & 0 & \beta_2 & 0 & 0 \\ 0 & 0 & \beta_0 & \beta_1 & \beta_2 & \beta_2 \\ \beta_2 & \beta_2 & \beta_1 & \beta_0 & 0 & 0 \\ 0 & 0 & \beta_2 & 0 & 0 & 0 \\ 0 & 0 & \beta_2 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \} \mathcal{S}_1 \\ \\ \\ \} \mathcal{S}_2 \end{matrix}$$

Figure 4.5: Structure of the error matrix $M_{\text{err}} = \mathbb{E}[T^o] - \mathbb{E}[T]$. Here, $\beta_0 = \mu_0^s - \mu_0$, $\beta_1 = \mu' - \mu^d$, and $\beta_2 = \mu^s - \mu^d$.

$(\mathcal{S}_1 \times \mathcal{S}_1) \backslash (\mathcal{S}_1 \cap \mathcal{S}_2 \times \mathcal{S}_1 \cap \mathcal{S}_2)$, then

$$\mathbb{E}[T_{uv}^o] = \mathbb{E}_Z \left[\frac{1}{n} \left(\sum_{j \in I_1} \nu \gamma_1^s + \lambda_0^2 (\gamma_2^s + \gamma_3^s) + \sum_{j \in I_2} \nu \gamma_1^d + \lambda_0^2 (\gamma_2^d + \gamma_3^d) \right) \right] \tag{4.68}$$

$$= \mu^s, \tag{4.69}$$

using $\mathbb{E}_Z[|I_1|] = \mathbb{E}_Z[|I_2|] = n/2$ and the definition in (4.63). The same result holds when $(u, v) \in (\mathcal{S}_2 \times \mathcal{S}_2) \backslash (\mathcal{S}_1 \cap \mathcal{S}_2 \times \mathcal{S}_1 \cap \mathcal{S}_2)$. On the other hand, when $(u, v) \in (\mathcal{S}_1 \backslash \mathcal{S}_2) \times (\mathcal{S}_2 \backslash \mathcal{S}_1)$,

$$\mathbb{E}[T_{uv}^o] = \mathbb{E}_Z \left[\frac{1}{n} \left(\sum_{j \in I_1} \nu \gamma_1^d + \lambda_0^2 (\gamma_2^d + \gamma_3^d) + \sum_{j \in I_2} \nu \gamma_1^d + \lambda_0^2 (\gamma_2^d + \gamma_3^d) \right) \right] \tag{4.70}$$

$$= \mu^d. \tag{4.71}$$

The same result holds when $(u, v) \in (\mathcal{S}_2 \backslash \mathcal{S}_1) \times (\mathcal{S}_1 \backslash \mathcal{S}_2)$. Finally, $(u, v) \in (\mathcal{S}_1 \cap \mathcal{S}_2) \times (\mathcal{S}_1 \cap \mathcal{S}_2)$,

$$\mathbb{E}[T_{uv}^o] = \mathbb{E}_Z \left[\frac{1}{n} \left(\sum_{j \in I_1} \nu \gamma_1^s + \lambda_0^2 (\gamma_2^s + \gamma_3^s) + \sum_{j \in I_2} \nu \gamma_1^s + \lambda_0^2 (\gamma_2^s + \gamma_3^s) \right) \right] \tag{4.72}$$

$$= 2\mu^s - \mu^d \stackrel{\text{def}}{=} \mu'. \tag{4.73}$$

This structure is depicted in Figure 4.4. Our next objective is to study the spectrum

of this matrix, and to bound its spectral norm and eigengap. We will do so by first expressing $\mathbb{E}[T^o]$ as

$$\mathbb{E}[T^o] = \mathbb{E}[T] + M_{\text{err}} \tag{4.74}$$

where M_{err} represents the error matrix $\mathbb{E}[T^o] - \mathbb{E}[T]$. The spectra of $\mathbb{E}[T^o]$ and $\mathbb{E}[T]$ can be related using Weyl's inequality. In particular, for every $i \in [k_{\text{un}}]$, it holds that

$$|\nu_i^o - \nu_i| \leq \|M_{\text{err}}\|_{op}, \tag{4.75}$$

where ν_i and ν_i^o represent the ith largest eigenvalues of $\mathbb{E}[T]$ and $\mathbb{E}[T^o]$, respectively. Thus, by triangle inequality, we have

$$\|\mathbb{E}[T^o]\|_{op} \leq \|\mathbb{E}[T]\|_{op} + \|M_{\text{err}}\|_{op}, \tag{4.76}$$

and using the relation in (4.75) we get

$$\begin{aligned}
\Delta_{\min}^o &= \min\{\nu_1^o - \nu_2^o,\ \nu_2^o - \nu_3^o\} && (4.77)\\
&\geq \min\Big\{\nu_1 - \|M_{\text{err}}\|_{op} - (\nu_2 + \|M_{\text{err}}\|_{op}), \nu_2 - \|M_{\text{err}}\|_{op} - (\nu_3 + \|M_{\text{err}}\|_{op})\Big\} && (4.78)\\
&= \Delta_{\min} - 2\|M_{\text{err}}\|_{op}. && (4.79)
\end{aligned}$$

It can be seen from the representation in Figure 4.5 that $M_{\text{err}} = \mathbb{E}[T^o] - \mathbb{E}[T]$ will be a sparse matrix. In particular, the diagonal is also sparse with non zeros placed only along indices in $\mathcal{S}_1 \cap \mathcal{S}_2$ (with value $\mu_0^s - \mu_0$). Letting $k_{\text{int}} = |\mathcal{S}_1 \cap \mathcal{S}_2|$, we have

$$\begin{aligned}
\|M_{\text{err}}\|_{op} \leq \texttt{Tr}(M_{\text{err}}) &= k_{\text{int}}(\mu_0^s - \mu_0) && (4.80)\\
&= k_{\text{int}}\left(\mu_0^s - \frac{1}{2}(\mu_0^s + \mu_0^d)\right) && (4.81)\\
&\leq k_{\text{int}}(\mu_0^s - \mu_0^d). && (4.82)
\end{aligned}$$

Using the definitions of μ_0^s and μ_0^d from (4.51) and (4.52), for arbitrary $u \neq i \neq j$,

$$\mu_0^s - \mu_0^d = \nu\mathbb{E}\left[\|\Phi_u\|_2^8 - (\Phi_u^\top \Phi_i)^4\right] \tag{4.83}$$

$$+ 2\lambda_0^2\mathbb{E}\left[2\|\Phi_u\|_2^4 \sum_{i\in S\setminus u} (\Phi_u^\top\Phi_i)^2 - (\Phi_u^\top\Phi_i)^2(\Phi_u^\top\Phi_j)^2\right] \tag{4.84}$$

$$\leq \nu c - \nu\left(\frac{2}{m^2} + \frac{1}{m^3}(c_2^2 - 2)\right) \tag{4.85}$$

$$+ 2\lambda_0^2\frac{2k}{m}\left(1 + \frac{3}{m}(c_2 - 1) + \frac{1}{m^2}(c_3 - 3c_2 + 2)\right) - \frac{4\lambda_0^2}{m^2}\left(1 + \frac{1}{m}(c_2 - 1)\right) \tag{4.86}$$

$$\leq c\left(\nu + \lambda_0^2\frac{k}{m}\right), \tag{4.87}$$

where $c > 1$ is an absolute constant. Plugging this into (4.79) gives

$$\Delta_{\min}^o \geq \Delta_{\min} - ck_{int}\left(\nu + \lambda_0^2\frac{k}{m}\right) \tag{4.88}$$

$$\geq c\lambda_0^2\left(k - k_{\text{int}}\frac{k}{m}\right) \tag{4.89}$$

$$= c\lambda_0^2 k\left(1 - \frac{k_{\text{int}}}{m}\right), \tag{4.90}$$

where we used (4.66) and omitted the dependence on the moments λ_0 and ν for simplicity. Assuming $k_{\text{int}}/m \leq c'$ for some $c' < 1$, we have

$$\Delta_{\min}^o \geq c''\lambda_0^2 k. \tag{4.91}$$

Finally, from (4.76) and (4.82), we get

$$\|\mathbb{E}\left[T^o\right]\|_{op} \leq c\lambda_0^2\left(\frac{k^3}{m^2} + k_{int}\frac{k}{m}\right) \leq c\lambda_0^2\frac{k^3}{m^2}. \tag{4.92}$$

Theorem 4.8.1 (Davis-Kahan)**.** *Let A and $\hat{A}$ be $d \times d$ symmetric matrices with eigenvalues $\nu_1 \geq \cdots \geq \nu_d$ and $\hat{\nu}_1 \geq \cdots \geq \hat{\nu}_d$, respectively. For a fixed $i \in [d]$, let $\Delta_{\min} \stackrel{\text{def}}{=} \min\{\nu_{i-1} - \nu_i, \nu_i - \nu_{i+1}\} > 0$ be the eigengap around the ith largest eigenvalue of*

A. Let v_i and $\hat{v}_i$ be the normalized eigenvectors corresponding to the ith largest eigenvalues of A and B, respectively. Then,

$$\sin(\angle(v_i, \hat{v}_i)) \leq \frac{2\|A - \hat{A}\|_{op}}{\Delta_{\min}}. \tag{4.93}$$

4.9 Useful lemmas

Lemma 4.9.1. *Let X be a subexponential random variable with parameters v^2 and $b > 0$, i.e., for every $t > 0$,*

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

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

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

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, [13, Theorem 2.3]) to derive the moment bound. We have,

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

that is,

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

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

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

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

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

which after simplification yields

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

□

Lemma 4.9.2. *Let $X \in \mathbb{R}^d$ be a mean zero random vector with independent entries such that $\mathbb{E}[X_i^2] = \lambda_0$ and $\mathbb{E}[X_i^4] = \rho$ for all $i \in [d]$. Then, for every $a, b \in \mathbb{R}^d$,*

$$\mathbb{E}\left[(X^\top ab^\top X)^2\right] = \rho \sum_{i=1}^d a_i^2 b_i^2 + \lambda_0^2 \sum_{i \neq j} (a_i^2 b_j^2 + a_i b_i a_j b_j).$$

In particular,

$$\mathbb{E}\left[(X^\top aa^\top X)^2\right] = \rho \sum_{i=1}^d a_i^4 + 2\lambda_0^2 \sum_{i \neq j} a_i^2 a_j^2.$$

Remark 4.9.3. *If the second and fourth moments are related as $\rho = 2\lambda_0^2 = 2c$ for some absolute constant c, then the result simplifies to $\mathbb{E}\left[(X^\top ab^\top X)^2\right] = c((a^\top b)^2 + \|a\|_2^2\|b\|_2^2)$.*

Proof. To start with, we note that the quadratic form $X^\top ab^\top X$ is a subexponential random variable since X is subgaussian. Although this fact can be used to derive upper bounds on the moments of $X^\top ab^\top X$, we would like to explicitly compute the second moment. We have,

$$\begin{aligned}
\mathbb{E}\left[(X^\top ab^\top X)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^d a_i b_i X_i^2 + \sum_{i \neq j} a_i b_j X_i X_j\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^d a_i b_i X_i^2\right)^2 + \left(\sum_{i \neq j} a_i b_j X_i X_j\right)^2 + 2\sum_{i=1}^d a_i b_i X_i^2 \sum_{i \neq j} a_i b_j X_i X_j\right]
\end{aligned}$$

$$= \mathbb{E}\left[\sum_{i=1}^{d} a_i^2 b_i^2 X_i^4 + \sum_{i\neq j} a_i b_i a_j b_j X_i^2 X_j^2 + \sum_{i\neq j} a_i^2 b_j^2 X_i^2 X_j^2\right].$$

Using $\mathbb{E}[X_i^2] = \lambda_0$ and $\mathbb{E}[X_i^4] = \rho$, we get

$$\mathbb{E}\left[(X^\top ab^\top X)^2\right] = \rho \sum_{i=1}^{d} a_i^2 b_i^2 + \lambda_0^2 \sum_{i\neq j} (a_i^2 b_j^2 + a_i b_i a_j b_j).$$

□

Lemma 4.9.4. *Let X, Y, Z and W be independent random vectors taking values in $\mathbb{R}^m$, with independent entries that are zero mean with variance $1/m$. Additionally, for every $i \in [m]$, let $\mathbb{E}\left[Z_i^{2q}\right] = c_q/m^q$, for q=2, 3, 4 and a constant c_q that depends only on q. Then, the following results hold:*

(i) $\mathbb{E}[\|Z\|_2^4] = 1 + \frac{1}{m}(c_2 - 1)$

(ii) $\mathbb{E}[\|Z\|_2^6] = 1 + \frac{3}{m}(c_2 - 1) + \frac{1}{m^2}(c_3 - 3c_2 + 2)$

(iii) $\mathbb{E}[\|Z\|_2^8] = 1 + \frac{6}{m}(c_2 - 1) + \frac{1}{m^2}(11 - 18c_2 + 6c_2^2 + 4c_3) + \frac{1}{m^3}(c_4 - 4c_3 - 6c_2^2 + 12c_2 - 6)$

(iv) $\mathbb{E}\left[(X^\top Y)^4\right] = \frac{2}{m^2} + \frac{1}{m^3}(c_2^2 - 2)$

(v) $\mathbb{E}\left[\|Z\|_2^4 (Z^\top W)^2\right] = \frac{1}{m}\left(1 + \frac{3}{m}(c_2 - 1) + \frac{1}{m^2}(c_3 - 3c_2 + 2)\right)$

(vi) $\mathbb{E}\left[(X^\top Z)^2 (X^\top W)^2\right] = \frac{1}{m^2}\left(1 + \frac{1}{m}(c_2 - 1)\right)$

(vii) $\mathbb{E}\left[\|Z\|_2^2 \|W\|_2^2 (Z^\top W)^2\right] = \frac{1}{m}\left(1 + \frac{1}{m}(c_2 - 1)\right)^2$

(viii) $\mathbb{E}\left[\|Z\|_2^2 (W^\top Z)(X^\top Z)(X^\top W)\right] = \frac{1}{m^2}\left(1 + \frac{1}{m}(c_2 - 1)\right)$

(ix) $\mathbb{E}\left[(Z^\top X)(Z^\top Y)(W^\top X)(W^\top Y)\right] = \frac{1}{m^3}$

(x) $\mathbb{E}\left[(X^\top Y)^2\right] = \frac{1}{m}$.

Proof. (i)

$$\mathbb{E}\left[\|Z\|_2^4\right] = \mathbb{E}\left[\sum_{i=1}^m Z_i^4 + \sum_{i\neq j} Z_i^2 Z_j^2\right]$$
$$= \frac{c_2}{m} + \frac{m-1}{m} = 1 + \frac{1}{m}(c_2 - 1).$$

(ii)

$$\mathbb{E}\left[\|Z\|^6\right] = \mathbb{E}\left[(Z_1^2 + \ldots + Z_m^2)^2(Z_1^2 + \ldots + Z_m^2)\right]$$
$$= \mathbb{E}\left[\left(\sum_{i=1}^m Z_i^4 + \sum_{i\neq j} Z_i^2 Z_j^2\right)\left(\sum_{t=1}^m Z_t^2\right)\right]$$
$$= \mathbb{E}\left[\sum_{i=1}^m Z_i^4 \sum_{t=1}^m Z_t^2 + \sum_{t=1}^m Z_t^2 \sum_{i\neq j} Z_i^2 Z_j^2\right].$$

For the first term,

$$\mathbb{E}\left[\sum_{i=1}^m Z_i^4 \sum_{t=1}^m Z_t^2\right] = \mathbb{E}\left[\sum_{i=1}^m Z_i^6 + \sum_{i\neq t} Z_i^4 Z_t^2\right]$$
$$= m\frac{c_3}{m^3} + m(m-1)\frac{c_2}{m^2}\frac{1}{m} = \frac{1}{m^2}(c_3 - c_2) + \frac{c_2}{m}, \tag{4.94}$$

and for the second term,

$$\mathbb{E}\left[\sum_{t=1}^m Z_t^2 \sum_{i\neq j} Z_i^2 Z_j^2\right] = \mathbb{E}\left[2\sum_{t\neq i} Z_t^4 Z_i^2 + \sum_{t\neq i\neq j} Z_t^2 Z_i^2 Z_j^2\right]$$
$$= 2m(m-1)\frac{c_2}{m^2}\frac{1}{m} + m(m-1)(m-2)\frac{1}{m^3}$$
$$= 1 + \frac{1}{m}(2c_2 - 3) - \frac{2}{m^2}(c_2 - 1)$$

Thus,

$$\mathbb{E}\left[\|Z\|^6\right] = 1 + \frac{3}{m}(c_2 - 1) + \frac{1}{m^2}(c_3 - 3c_2 + 2).$$

(iii)

$$\begin{aligned}\mathbb{E}\left[\|Z\|^8\right] =&\mathbb{E}\left[(Z_1^2+\cdots+Z_m^2)^4\right]\\ =&m\mathbb{E}\left[Z_1^8\right]+\binom{m}{2}\frac{4!}{3!}2\mathbb{E}\left[Z_1^6Z_2^2\right]+\binom{m}{2}\frac{4!}{2!2!}2\mathbb{E}\left[Z_1^4Z_2^4\right]\\ &+\binom{m}{3}\frac{4!}{2!}3\mathbb{E}\left[Z_1^4Z_2^2Z_3^2\right]+\binom{m}{4}4!\mathbb{E}\left[Z_1^2Z_2^2Z_3^2Z_4^2\right]\\ =&1+\frac{6}{m}(c_2-1)+\frac{1}{m^2}(11-18c_2+6c_2^2+4c_3)\\ &+\frac{1}{m^3}(c_4-4c_3-6c_2^2+12c_2-6).\end{aligned}$$

(iv) To compute $\mathbb{E}\left[(X^\top Y)^4\right]$, we first note that

$$\begin{aligned}\mathbb{E}\left[(X^\top Y)^4|X\right] &= \mathbb{E}\left[(Y^\top XX^\top Y)^2|X\right]\\ &= \mathbb{E}\left[Y_1^4\right]\sum_{i=1}^m X_i^4+2(\mathbb{E}\left[Y_1^2\right])^2\sum_{i\neq j}X_i^2X_j^2\\ &= \frac{c_2}{m^2}\sum_{i=1}^m X_i^4+2\left(\frac{1}{m}\right)^2\sum_{i\neq j}X_i^2X_j^2,\end{aligned}$$

where we used Lemma 4.9.2 in the second step. This gives

$$\begin{aligned}\mathbb{E}\left[(X^\top Y)^4\right] &= \frac{c_2}{m}\mathbb{E}\left[X_1^4\right]+\frac{2(m-1)}{m}(\mathbb{E}\left[X_1^2\right])^2\\ &= \frac{c_2^2}{m^3}+\frac{2(m-1)}{m^3}=\frac{2}{m^2}+\frac{1}{m^3}(c_2^2-2).\end{aligned}$$

(v) Similar to the previous calculation, we first compute the conditional expectation to get

$$\mathbb{E}\left[\|Z\|_2^4(Z^\top W)^2|Z\right]=\|Z\|_2^4\left(\sum_{i=1}^m\mathbb{E}\left[Z_i^2W_i^2|Z\right]+\sum_{i\neq j}\mathbb{E}\left[Z_iW_iZ_jW_j|Z\right]\right)=\|Z\|_2^4\frac{\|Z\|_2^2}{m},$$

which gives

$$\mathbb{E}\left[\|Z\|_2^4(Z^\top W)^2\right]=\frac{1}{m}\mathbb{E}\left[\|Z\|_2^6\right]=\frac{1}{m}\left(1+\frac{3}{m}(c_2-1)+\frac{1}{m^2}(c_3-3c_2+2)\right).$$

(vi) We have

$$\mathbb{E}\left[(X^\top Z)^2(X^\top W)^2|X\right] = \mathbb{E}\left[(X^\top Z)^2|X\right]\mathbb{E}\left[(X^\top W)^2|X\right] = \frac{\|X\|_2^2}{m}\cdot\frac{\|X\|_2^2}{m}.$$

Thus,

$$\mathbb{E}\left[(X^\top Z)^2(X^\top W)^2\right] = \frac{1}{m^2}\bigg(1+\frac{1}{m}(c_2-1)\bigg).$$

(vii)

$$\begin{aligned}
\mathbb{E}\left[\|Z\|_2^2\|W\|_2^2(Z^\top W)^2|Z\right] =& \|Z\|_2^2\,\mathbb{E}\left[\|W\|_2^2(Z^\top W)^2|Z\right] \\
=& \|Z\|_2^2\bigg(\sum_{i=1}^m \mathbb{E}\left[\|W\|_2^2 Z_i^2 W_i^2|Z\right] + \sum_{i\neq j}\mathbb{E}\left[\|W\|_2^2 W_iW_jZ_iZ_j|Z\right]\bigg) \\
=& \|Z\|_2^2\sum_{i=1}^m Z_i^2\mathbb{E}\left[W_i^4 + \sum_{l\neq i} W_i^2W_l^2\right] \\
&+ \|Z\|_2^2\sum_{i\neq j} Z_iZ_j\mathbb{E}\left[W_i^3W_j + W_j^3W_i + \sum_{l\neq i,\ l\neq j} W_l^2W_iW_j\right] \\
=& \|Z\|_2^2\sum_{i=1}^m Z_i^2\bigg(\frac{c_2}{m^2}+\frac{m-1}{m^2}\bigg) = \|Z\|_2^4\bigg(\frac{1}{m}+\frac{c_2-1}{m^2}\bigg).
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\|Z\|_2^2\|W\|_2^2(Z^\top W)^2\right] = \frac{1}{m}\bigg(1+\frac{c_2-1}{m}\bigg)^2.$$

(viii)

$$\begin{aligned}
\mathbb{E}\left[\|Z\|_2^2(W^\top Z)(X^\top Z)(X^\top W)|Z,W\right] &= \|Z\|_2^2(W^\top Z)\mathbb{E}\left[X^\top WZ^\top X|W,Z\right] \\
&= \|Z\|_2^2(W^\top Z)\frac{Z^\top W}{m}.
\end{aligned}$$

Using similar arguments as in the proof of (v),

$$\mathbb{E}\left[\|Z\|_2^2(W^\top Z)(X^\top Z)(X^\top W)\right] = \frac{1}{m^2}\left(1+\frac{c_2-1}{m}\right).$$

(ix)

$$\begin{aligned}\mathbb{E}\left[(Z^\top X)(Z^\top Y)(W^\top X)(W^\top Y)|X,Y,W\right] &= (W^\top X)(W^\top Y)\mathbb{E}\left[Z^\top XY^\top Z|X,Y\right] \\ &= (W^\top X)(W^\top Y)\frac{X^\top Y}{m}\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[(Z^\top X)(Z^\top Y)(W^\top X)(W^\top Y)\right] &= \frac{1}{m}\mathbb{E}_{X,Y}\left[\mathbb{E}_W\left[(W^\top X)(W^\top Y)(X^\top Y)|X,Y\right]\right] \\ &= \frac{1}{m}\mathbb{E}_{X,Y}\left[(X^\top Y)\mathbb{E}_W\left[W^\top XY^\top W|X,Y\right]\right] \\ &= \frac{1}{m^2}\mathbb{E}_{X,Y}\left[(X^\top Y)^2\right] = \frac{1}{m^3}.\end{aligned}$$

(x)

$$\mathbb{E}\left[(X^\top Y)^2\right] = \sum_{i=1}^m \mathbb{E}\left[X_i^2Y_i^2\right] + \sum_{i\neq j}\mathbb{E}\left[X_iY_iX_jY_j\right] = \frac{1}{m}.$$

□

Chapter 5

Conclusions

5.1 Summary

We studied the problem of support recovery from linear measurements under the constraint that we can only obtain very few measurements per sample. For the case of a single unknown support, we derived tight upper and lower bounds on the sample complexity and saw that the measurement constraint leads to an increase in the sample complexity compared to the measurement-rich regime. Our upper bound results under both random and deterministic inputs showed that a simple variance estimation based procedure achieves the optimal scaling when the measurement matrices satisfy a separation condition. In summary, our results showed a change in the sample complexity of this problem as we move from the measurement-rich $m > k$ regime to the measurement-constrained $m < k$ regime. We then studied the case of multiple supports under similar measurement constraints. We used a combination of the variance estimation step and a spectral clustering step to estimate the underlying supports, and provided an upper bound on the sample complexity under a mixture model prior on the inputs.

5.2 Directions for further work

We outline some interesting directions that can be studied in the context of the work presented in this thesis.

1. Multiple support recovery with arbitrary overlaps

   Our current result for multiple support recovery in the general ℓ case is for disjoint supports in a measurement-constrained setting. It would be interesting to extend the algorithm to handle overlapping supports. While some simple heuristics work in practice, we are not aware of any theoretical results for this setting when $m < k$. In particular, extending the ℓ-means step to handle overlaps and characterizing the spectrum of $\mathbb{E}[T]$ with arbitrary overlaps are both challenging in general.

2. Subspace recovery under measurement constraints

   Another direction that can be considered is when the samples are sparse in an *unknown* basis, namely when the samples are drawn from a union of subspaces, and we are given access to very few measurements from each sample. Such data can be modeled using a mixture of degenerate Gaussians with the component Gaussians having a low-rank covariance matrix, and similar to our setting in the sparse case, one could consider designing algorithms for recovering the unknown subspaces. The question of labeling the samples first (as opposed to estimating the subspaces first) is also an interesting question which has been looked at [85], [74] although these algorithms are not designed for the measurement-constrained setting.

3. Lower bound for multiple support recovery

   A lower bound on the sample complexity of the multiple support recovery problem is not known in the $m < k$ regime. A key challenge here is to characterize the distance between mixture distributions, and using a component-wise bound does not yield tight results. In particular, using an approach similar to the common support case, we can model the inputs as being drawn from a Gaussian mixture with components that have zero mean and sparse, diagonal covariance matrices. The KL divergence between pairs of output distributions cannot be expressed in closed form, and relaxing

it to component-wise distances leads roughly to the same result as in the $\ell = 1$ case. One could use other distance measures, however obtaining tight bounds under this covariance-based prior on the inputs is difficult.

4. Moment-based estimators for measurement-constrained settings

   Both our estimators are based on the idea that when the gram matrix of the measurement matrices roughly behaves like the identity matrix, the measurements can be "inverted" to get proxy samples, and sample averages of their higher moments can be used to find interesting structure in the data. This approach is able to work with very few measurements per sample. A more general understanding of this procedure can help in designing estimators for other problems in measurement-constrained settings.

Bibliography

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

[2] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints I: Lower bounds from chi-square contraction," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7835–7855, 2020.

[3] 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.

[4] 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.

[5] 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.

[6] 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

[7] 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.

[8] 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.

[9] 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.

[10] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *The Annals of Statistics*, vol. 41, no. 4, pp. 1780 – 1815, 2013.

[11] L. Birgé, *An alternative point of view on Lepski's method*, ser. Lecture Notes–Monograph Series. Beachwood, OH: Institute of Mathematical Statistics, 2001, vol. Volume 36, pp. 113–133.

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

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

[14] T. T. Cai, Z. Ren, and H. H. Zhou, "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1 – 59, 2016.

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

[16] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 267–288, Dec. 2008.

[17] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

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

[19] 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

[20] 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.

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

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

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

[24] Y. Deshpande and A. Montanari, "Sparse PCA via covariance thresholding," *J. Mach. Learn. Res.*, vol. 17, pp. 141:1–141:41, 2016.

[25] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.*, 2005, pp. 1537–1541.

[26] 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.

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

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

[29] S. Foucart and D. Koslicki, "Sparse recovery by means of nonnegative least squares," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 498–502, 2014.

[30] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Boston: Birkhaüser, 2013.

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

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] S. Khanna and C. R. Murthy, "Rényi divergence based covariance matching pursuit of joint sparse support," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications, SPAWC 2017, Sapporo, Japan, July 3-6, 2017*, 2017, pp. 1–5.

[38] C. G. Khatri and C. R. Rao, "Solutions to some functional equations and their applications to characterization of probability distributions," *Sankhya, Series A*, vol. 30, pp. 167–180, 1968.

[39] 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.

[40] A. Koochakzadeh and P. Pal, "A greedy approach for correlation-aware sparse support recovery," in *Compressive Sensing VII: From Diverse Modalities to Big Data Analytics*, vol. 10658. SPIE, 2018, pp. 89 – 95.

[41] 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.

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

[43] 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.

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

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

[46] 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.

[47] M. Ndaoud and A. B. Tsybakov, "Optimal variable selection and adaptive noisy compressed sensing," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2517–2532, 2020.

[48] S. N. Negahban and M. J. Wainwright, "Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3841–3863, 2011.

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

[50] 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.

[51] 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.

[52] R. Oliveira, "Sums of random Hermitian matrices and an inequality by rudelson," *Electron. Commun. Probab.*, vol. 15, pp. 203–212, 2010.

[53] P. Pal and P. P. Vaidyanathan, "IEEE international conference on acoustics, speech and signal processing, ICASSP 2014, florence, italy, may 4-9, 2014," 2014, pp. 1851–1855.

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

[55] 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.

[56] I. F. Pinelis and S. A. Utev, "Estimates of the moments of sums of independent random variables," *Theory of Probability & Its Applications*, vol. 29, no. 3, pp. 574–577, 1985. [Online]. Available: https://doi.org/10.1137/1129075

[57] 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

[58] 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.

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

[60] 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

[61] ——, "Multiple support recovery using very few measurements per sample," 2021. [Online]. Available: https://github.com/lekshmi-ramesh/SupportRecovery/blob/master/LR-CM-HT-21b.pdf

[62] ——, "Phase transition for support recovery from gaussian linear measurements," 2021. [Online]. Available: https://arxiv.org/abs/2102.00235

[63] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6976–6994, 2011.

[64] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *IEEE International Symposium on Information Theory*, 2008, pp. 2187–2191.

[65] 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.

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

[67] M. Rudelson and R. Vershynin, "Sampling from large matrices: An approach through geometric functional analysis," *J. ACM*, vol. 54, no. 4, p. 21, 2007.

[68] ——, "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

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

[70] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," in *Advances in Neural Information Processing Systems*, 2002, pp. 1041–1048.

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

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

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

[74] P. A. Traganitis and G. B. Giannakis, "Sketched subspace clustering," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1663–1675, 2018.

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

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

[77] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *IEEE ICASSP, Philadelphia, Pennsylvania, USA, March*, 2005, pp. 721–724.

[78] ——, "Simultaneous sparse approximation via greedy pursuit," in *IEEE ICASSP, Philadelphia, Pennsylvania, USA, March*, 2005, pp. 721–724.

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

[80] 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.

[81] 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.

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

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

[84] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2967–2979, 2010.

[85] Y. Wang, Y.-X. Wang, and A. Singh, "A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1422–1431. [Online]. Available: http://proceedings.mlr.press/v37/wange15.html

[86] 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.

[87] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[88] ——, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7-2, pp. 3704–3716, 2007.

[89] 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.

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

[91] I. Zadik and D. Gamarnik, "Sparse high-dimensional linear regression. algorithmic barriers and a local search algorithm," November 2017. [Online]. Available: https://arxiv.org/abs/1711.04952

[92] 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

[93] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, p. 2006, 2004.