Electronic Journal of Statistics
Vol. 11 (2017) 2647–2678
ISSN: 1935-7524
DOI: 10.1214/17-EJS1267

Recovering block-structured activations using compressive measurements

Sivaraman Balakrishnan

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: siva@stat.cmu.edu

Mladen Kolar

Booth School of Business, University of Chicago, Chicago, IL 60637, USA
e-mail: mkolar@chicagobooth.edu

Alessandro Rinaldo

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: arinaldo@stat.cmu.edu

and

Aarti Singh

Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: aarti@cs.cmu.edu

Abstract: We consider the problems of detection and support recovery of a contiguous block of weak activation in a large matrix, from noisy, possibly adaptively chosen, compressive (linear) measurements. We precisely characterize the tradeoffs between the various problem dimensions, the signal strength and the number of measurements required to reliably detect and recover the support of the signal, both for passive and adaptive measurement schemes. In each case, we complement algorithmic results with information-theoretic lower bounds. Analogous to the situation in the closely related problem of noisy compressed sensing, we show that for detection neither adaptivity, nor structure reduce the minimax signal strength requirement. On the other hand we show the rather surprising result that, contrary to the situation in noisy compressed sensing, the signal strength requirement to recover the support of a contiguous block-structured signal is strongly influenced by both the signal structure and the ability to choose measurements adaptively.

MSC 2010 subject classifications: Primary 62F03; secondary 62F10.
Keywords and phrases: Adaptive sensing, linear measurements, structured normal means.

Received August 2016.

Contents

1. Introduction

The estimation of a sparse signal from noisy observations is a problem of central interest in statistics and signal processing. In this paper, we consider the problems of detecting the presence of and estimating the support of a small contiguous block of signal that is embedded in a large data matrix, given access to noisy compressive measurements, i.e., linear combinations of the matrix entries corrupted with noise. Such problems arise in a variety of applications, including remote sensing [23, 43], computational biology [47], image processing [14] and anomaly detection [4, 44].

Broadly, our work is part of a large body of literature on estimating a high-dimensional structured matrix from noisy measurements. This problem has received widespread attention in applications such as community detection, multiple linear regression, sparse PCA and ranking problems (see for instance the paper [18] and references therein). Our focus in this paper is on highly structured matrices for which a sparse set of consecutive rows and columns are active. Data matrices with sparse, contiguous block-structured signal components form an idealized model for signals arising in several real-world applications. For

instance, the expression pattern resulting from genes, grouped by pathways, expressed under the influence of drugs, grouped by similarity. The block structure in this case arises from the fact that groups of genes (belonging to common pathways, say) are co-expressed under the influence of sets of similar drugs [47].

As an illustrative example consider Figure 1. The figure considers an application of object detection and localization via compressive sensing. The illustrated application further shares similarities with other applications including search and rescue in open areas, and gas leak or radiation detection [32, 22]. We consider two problems in this context: the hypothesis testing problem of detecting the presence of an object, and the localization or support recovery problem of precisely locating the object in the matrix. Additionally, we compare and contrast two measurement paradigms: a passive measurement scheme where the measurement vectors are chosen non-adaptively and in advance of observing measurement outcomes, and an adaptive or sequential scheme where each measurement vector can be chosen after observing the outcome of previous measurements.

1.1. Our contributions

Our primary contributions are to establish the fundamental limits for adaptive and passive measurement schemes for detection and localization of a contiguous block of positive activation. In this direction, our first contribution (Theorem 1) establishes the fundamental limits for detection of a contiguous block of activation. Similar to situation of detecting a sparse vector, as studied by Arias-Castro [2], we show that neither structure, nor the ability to choose measurements sequentially, help in the detection problem. In this setting, analogous to the sparse vector setting, a fairly naive passive scheme is optimal. Our proof technique follows that of Arias-Castro [2]: however, we require a slightly more refined analysis here in order to show that the lower bound on adaptive schemes continues to hold despite the additional signal structure.

Our second contribution, is to determine upper and lower bounds for the problem of localization from passive measurements. In the sparse vector setting this problem has been considered in a variety of papers with fundamental limits, for Gaussian measurement ensembles, appearing in the work of Wainwright [45]. In Theorem 2, we establish analogous results for the case of a contiguous block of activation. Our lower bounds follow from the construction of two packing sets, that respectively capture the difficulty of approximately and exactly localizing the support of the signal, combined with classical information theoretic techniques involving Fano's inequality. For our upper bound we build on the techniques of Wainwright [45], refining them to account for the much lower cardinality of the class of signals under consideration in this paper.

Our third contribution is to establish upper and lower bounds for the problem of localization from adaptive, sequentially chosen measurements. In the sparse vector case upper and lower bounds for adaptive measurement schemes follow from a sequence of past work [3, 34, 33, 19, 15]. We use techniques of Arias-Castro [2] in order to develop modifications to the classical Fano argument to

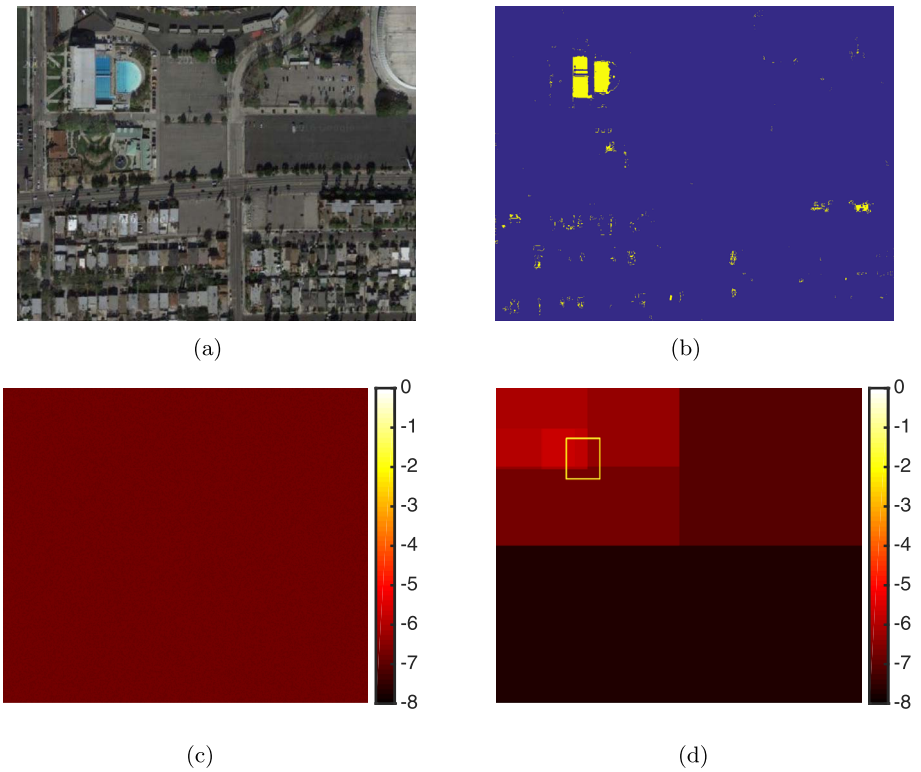

FIG 1. *An illustrative application of object detection from Landsat images. Figure (a) shows the original RGB image, and Figure (b) shows the de-noised blue channel of this image. The goal is to detect/localize the object (swimming pool) in the image, via a small number of compressed measurements. Figures (c) and (d) show the (log) energy profile of a passive and sequential sensing scheme respectively. In this context, adaptive sensing can be accomplished via a compressive imaging camera whose resolution or field of view is adjusted to zoom in on specific areas. The passive sensing scheme uses roughly uniform energy on each pixel while the adaptive scheme adaptively allocates its sensing budget, and spends a considerable amount of sensing energy in precisely localizing the signal.*

account for adaptive measurements: once again we adopt multiple constructions to capture the difficulty of approximately and precisely localizing the block of activation. For our upper bounds we build on the work of Malloy and Nowak [33, 34] (see also [19]), and analyze a compressive binary search algorithm. At a high-level, as illustrated in Figure 1, the compressive binary search algorithm sequentially invests its sensing energy on quadrants that are likely to contain the block of activation. This coarse localization is then followed by a novel boundary detection algorithm to exactly identify the signal support. We provide an analysis of this algorithm together with an information-theoretic lower bound in Theorem 3.

These results taken together give a sharp characterization of the fundamental limits for detection and localization of a contiguous block of signal from passive

and adaptive compressive measurements. At a conceptual level while previous results for the sparse vector case indicate that adaptivity offers no improvement in a minimax sense, our results indicate that for localizing a contiguous block adaptivity and structure play a key role and that significant improvements are possible by using adaptive measurement schemes. More broadly, our results suggest that adaptive measurements can offer significant gains in highly structured settings and that further developments may be possible (see the papers [16, 31, 39] for some recent progress) in this direction.

The rest of the paper is organized as follows: In Section 2 we set up the measurement model, formally introduce the problems of detection and support recovery, and discuss related work. In Section 3 we present our main results and discuss their consequences. Section 4 is devoted to the proofs our main theorems, with the remaining technical details deferred to the Appendix. In Section 5 we provide a variety of simulation results, and we conclude with a discussion of extensions in Section 6.

Notation: We use $[n]$ to denote the set $\{1, \ldots, n\}$. $\mathbb{I}\{\mathcal{E}\}$ denotes the 0/1 indicator function for the event $\mathcal{E}$. For a vector $a \in \mathbb{R}^d$, $\mathrm{supp}(a) := \{j \ : \ a_j \neq 0\}$ denotes its support (with an analogous definition for matrices $\Theta \in \mathbb{R}^{d_1 \times d_2}$), $\|a\|_q$, $q \in [1, \infty)$. The ℓ_q-norm is defined as $\|a\|_q = (\sum_{i \in [d]} |a_i|^q)^{1/q}$ with the standard extensions for $q \in \{0, \infty\}$. For a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$, we use the notation $\mathrm{vec}(\Theta)$ to denote the vector in $\mathbb{R}^{d_1 d_2}$ formed by stacking the columns of Θ. We denote the Frobenius norm of Θ by $|||\Theta|||_{\mathrm{fro}}$ and its operator norm by $|||\Theta|||_{\mathrm{op}}$. For two matrices Θ_1 and Θ_2 of identical dimensions, $\langle\langle \Theta_1, \ \Theta_2 \rangle\rangle$ denotes the trace inner product, i.e. $\langle\langle \Theta_1, \ \Theta_2 \rangle\rangle = \mathrm{trace}(\Theta_1^T \Theta_2)$.

2. Background and problem setup

In this section, we provide some background on our measurement model, the problems of detection and support recovery, and the adaptive and passive measurement schemes we consider in this paper. Finally, we conclude this section with a survey of related work.

2.1. The measurement model

Throughout, we will assume that we collect n noisy linear measurements, $\{y_1, \ldots, y_n\}$ of an unknown signal matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ using measurement matrices $\{X_1, \ldots, X_n\}$. The measurement outcomes y_i are related to the measurement matrices X_i via the linear model:

$$y_i = \langle\langle \Theta^*, \ X_i \rangle\rangle + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are Gaussian with mean zero and variance σ^2. Following the convention of the majority of the literature on compressive sensing we normalize the sensing matrices X_i to satisfy one of the following two conditions: 1) $|||X_i|||_{\mathrm{fro}}^2 \leq 1$ or 2) $\mathbb{E}|||X_i|||_{\mathrm{fro}}^2 \leq 1$, for every $i \in [n]$. This normalization ensures

that every measurement is made with the same amount of energy (possibly in expectation). Typically, the first condition is enforced when the measurements are non-random while the second condition is enforced for random measurements.

We are interested in a highly structured setting where the true matrix Θ^* contains a contiguous block of positive activation of size $k_1 \times k_2$. Formally, a contiguous block $B \in \mathcal{B}$ is a collection of indices:

$$\mathcal{B} = \left\{ I_r \times I_c \ : \ \begin{array}{l} I_r \text{ and } I_c \text{ are contiguous subsets of } [d_1] \text{ and } [d_2], \\ |I_r| = k_1, |I_c| = k_2 \end{array} \right\}. \quad (2.2)$$

We let B^* denote the contiguous block of activation in Θ^*, and we denote the minimum signal strength by $\theta_{\min}$:

$$\theta_{\min} := \min_{(i,j) \in B^*} \Theta^*_{ij} > 0.$$

We note in passing that this measurement model could be equivalently vectorized. We maintain the matrix representation of Equation (2.1) in order to emphasize that the support of the signal is a contiguous block. We focus on the case when the size of the contiguous block, i.e, k_1, k_2 are known, but indicate in Section 6 methods that adapt to these parameters. In Section 6 we also briefly consider the case when the signal has mixed sign. Finally, in order to concisely present our results we focus on the case when

$$k_1 \leq d_1/2 \quad \text{and} \quad k_2 \leq d_2/2. \quad (2.3)$$

Our results consider two types of measurements schemes: passive measurement schemes where the measurement matrices $\{X_1, \ldots, X_n\}$ are chosen a priori, possibly from a random ensemble, and adaptive measurement schemes which may be implemented in a sequential fashion. Formally, in an adaptive scheme each measurement matrix X_i is a possibly randomized function of the prior measurement matrices and their corresponding outcomes $(y_j, X_j)_{j \in [i-1]}$.

2.2. Detection

The detection problem consists of deciding whether a contiguous block of signal exists in Θ^*. Studying the detection problem has multiple benefits: typically we can detect the presence of a signal using significantly fewer measurements and the detection problem is thus a natural precursor to the problems of estimation and support recovery. Furthermore, in this paper we build on algorithms for detection, and use them for approximate localization.

The detection problem involves hypothesis testing with a composite alternative. As a preliminary, for a given value $\theta_{\min} > 0$ we define the class of matrices:

$$\mathbf{\Theta}(\theta_{\min}) = \left\{ \Theta : \exists \ B \in \mathcal{B}, \text{ such that } \operatorname{supp}(\Theta) = B, \min_{(i,j) \in B} \Theta_{ij} \geq \theta_{\min} \right\}.$$

With this definition in place, the detection problem is to reliably distinguish the following hypotheses:

$$\begin{aligned} H_0 &: \quad \Theta^* = 0_{d_1 \times d_2} \\ H_1 &: \quad \Theta^* \in \mathbf{\Theta}(\theta_{\min}). \end{aligned} \tag{2.4}$$

A test T is a function that takes $(y_i, X_i)_{i\in[n]}$ as an input and outputs either 1, if the null hypothesis H_0 is rejected, and 0 otherwise. We denote by $\mathbb{P}_0$ and $\mathbb{P}_{\Theta^*}$ the joint probability distributions of $(y_i, X_i)_{i\in[n]}$ under the null hypothesis and alternative hypothesis respectively. For any test T, we define its risk as

$$R^{\mathrm{det}}(T) := \mathbb{P}_0\left[T\big((y_i, X_i)_{i\in[n]}\big) = 1\right] + \sup_{\Theta^*\in\mathbf{\Theta}(\theta_{\min})} \mathbb{P}_{\Theta^*}\left[T\big((y_i, X_i)_{i\in[n]}\big) = 0\right]. \tag{2.5}$$

The risk $R^{\mathrm{det}}(T)$ measures the sum of type I and maximal type II errors over the set of alternatives. The overall difficulty of the detection problem is quantified by the minimax detection risk:

$$R^{\mathrm{det}} := \inf_T R^{\mathrm{det}}(T),$$

where the infimum is taken over all measurable test functions. In Section 3.1 (see Theorem 1) we characterize necessary and sufficient conditions for distinguishing H_0 and H_1 as a function of $\theta_{\min}$.

2.3. Support recovery

The problem of identifying the support of the signal Θ^* is that of determining the exact location of the non-zero elements of Θ^*. We refer to this problem as that of support recovery or localization. Formally, we let Ψ be an estimator of B^*, i.e., a function that takes $(y_i, X_i)_{i\in[n]}$ as input and outputs an element of $\mathcal{B}$. We define the risk of any such estimator as

$$R^{\mathrm{sup}}(\Psi) := \sup_{\Theta^*\in\mathbf{\Theta}(\theta_{\min})} \mathbb{P}_{\Theta^*}\left[\Psi\big((y_i, X_i)_{i\in[n]}\big) \neq \mathrm{supp}(\Theta^*)\right],$$

while the minimax support recovery risk is

$$R^{\mathrm{sup}} := \inf_\Psi R^{\mathrm{sup}}(\Psi), \tag{2.6}$$

where the infimum is taken over all estimators Ψ. Like in the detection task, the minimax risk specifies the minimal risk of any support identification procedure. We focus in this paper on exact support recovery, and defer a discussion of recovery in, for instance, the Hamming metric to Section 6.

2.4. Related work

We conclude this section by highlighting some related work. In past work several researchers have analyzed statistical limits and proposed computationally tractable algorithms for using compressive measurements for the detection

TABLE 1

Summary of known results for the sparse vector case. $\theta_{\min}$ is the minimum (absolute) signal amplitude over all non-zero coordinates, and σ is the standard deviation of the noise. We use C to denote universal constants independent of the problem parameters.

	Detection	Localization	
Passive	$\frac{\theta_{\min}}{\sigma} \geq C\sqrt{\frac{d}{nk^2}}$ for positive signals.	$n \geq Ck\log(d/k)$	Wainwright [45]
	$\frac{\theta_{\min}}{\sigma} \geq C\sqrt{\frac{d}{nk}}$ for arbitrary signals.	$\frac{\theta_{\min}}{\sigma} \geq C\sqrt{\frac{d\log d}{n}}$	
Adaptive	Arias-Castro [2]	$n \geq Ck\log(d/k)$	Arias-Castro et al. [3]
		$\frac{\theta_{\min}}{\sigma} \geq C\sqrt{\frac{d\log k}{n}}$	Malloy and Nowak [33]

[21, 26, 2, 5, 28], estimation [20, 12, 13] and support recovery [46, 45, 1, 36] of sparse vectors. While these previous works focused on passive measurements, more recent work has considered adaptive compressive measurements [11, 3, 25, 33, 34, 2]. In more detail, the work of Haupt et al. [25] considers the estimation of a sparse vector using a procedure called compressive distilled sensing while Arias-Castro et al. [3] provide complementary lower bounds. Arias-Castro [2] provides both upper and lower bounds for adaptive detection of sparse vectors while the papers of Malloy and Nowak [33, 34] address the problem of support recovery.

In order to facilitate a subsequent comparison with the results that we present in this work, Table 1 summarizes known results for the detection and support recovery of a sparse vector using passive and adaptive compressive measurements. In each case we only provide sufficient conditions, and the corresponding references detail the precise assumptions under which these results hold. In Table 1, we follow the standardization used throughout this paper (and various others on this topic), i.e., the length of the vector is d and the number of non-zero coordinates in the vector is k. The number of measurements is n and each measurement is assumed to have unit ℓ_2 norm (in expectation, if the measurement is random).

The prior work described so far has been focused on inference for sparse, but otherwise unstructured, data vectors. We also note in passing the significant literature on estimation of structured matrices [18]: particularly low-rank matrices [35, 30, 37, 24], and those with cluster structured activations [40, 9, 10, 8, 29]. In this paper we focus on the relatively unexplored problems of detection and support recovery for highly structured signals from compressive measurements. The works of Baraniuk et al. [7], Huang et al. [27] have analyzed estimation under different forms of structured sparsity in the vector setting, for example, when the non-zero locations in a data vector form non-overlapping or partially-overlapping groups. These works do not however do not address the problems of detection and support recovery and do not consider the adaptive setting. Castro [15] and Tánczos and Castro [41] study support recovery of unstructured and structured signals from adaptive measurements when the signal is observed directly, i.e., not from compressive measurements. Most closely related to our own work is the work of Soni and Haupt [38]. They consider the case of adaptive sensing of tree-structured signals. However, the class of tree-structured signals is

quite different from the class of signals we consider, and the gains from adaptive sensing for the tree-structured signals are relatively limited (to a log factor as in the unstructured vector setting).

Finally, since the initial posting of our paper [6], several papers have considered adaptive compressed sensing with different signal structures using similar techniques. Concretely, the work of Castro and Tánczos [16] investigates various combinatorial signal structures, Soni and Haupt [39] consider tree-structured signals, while Krishnamurthy et al. [31] consider graph-structured signals.

3. Main results

In this section we present our main results concerning the detection and support recovery of contiguous block structured signals from passive and adaptive measurements. In more detail, Theorem 1 gives upper and lower bounds on the minimax testing risk R^{det} defined in Equation (2.5) for both passive and adaptive measurements, Theorem 2 gives upper and lower bounds on the minimax support recovery risk R^{sup} defined in Equation (2.6) when using passive measurements. Finally, Theorem 3 gives upper and lower bounds on R^{sup} when using adaptive measurements.

3.1. Detection of contiguous blocks

In this section we consider the detection problem of distinguishing between the null and alternate hypotheses defined in Equation (2.4). In order to establish an upper bound we consider the testing procedure proposed by Arias-Castro [2]. The measurement vectors are chosen passively as

$$X_i = (d_1 d_2)^{-1/2} \mathbf{1}_{d_1} \mathbf{1}'_{d_2} \quad i \in \{1, \ldots, n\},$$

where $\mathbf{1}_d$ is used to denote the vector of ones in $\mathbb{R}^d$. With these measurement vectors, the test is described as:

$$T\big((y_i)_{i\in[n]}\big) = \mathbb{I}\Big\{ \sum_{i=1}^{n} y_i > \tau \Big\}, \tag{3.1}$$

where τ is a threshold whose value will be prescribed to appropriately balance the probability of Type I and Type II errors. With the testing procedure in place, we have the following result that applies to measurements that may be chosen either passively or adaptively.

Theorem 1. *Fix any* $0 < \alpha < 1$.

1. *If*

$$\frac{\theta_{\min}}{\sigma} \leq 8(1-\alpha)\sqrt{\frac{d_1 d_2}{n k_1^2 k_2^2}},$$

then the minimax detection risk, based on n possibly adaptive measurements, is: $R^{\text{det}} \geq \alpha$.

2. *Conversely, there is a universal constant $C > 0$ such that if,*

$$\frac{\theta_{\min}}{\sigma} \geq C\sqrt{\frac{d_1 d_2}{nk_1^2 k_2^2}\log\left(\frac{1}{\alpha}\right)},$$

then the testing procedure described in Equation (3.1) with $\tau = \sigma\sqrt{2n\log(\alpha^{-1})}$ has $R^{\det}(T) \leq \alpha$.

Remarks:

- The proof of this theorem is given in Section 4. The upper bound follows essentially verbatim from the result of Arias-Castro [2]. The lower bound requires careful modification of the classical Fano argument in order to allow for adaptive measurements. Once again we follow the basic recipe outlined in [2], but we use a more refined analysis in order to show that the lower bound continues to hold despite the additional signal structure.
- It is worth noting that the contiguous block structure of the activation pattern does not play any role in the minimax detection problem. Indeed the rate established matches the known bounds for detection in the unstructured vector case (see Table 1). We will contrast this to the problem of support recovery below. Furthermore, the procedure that achieves the adaptive lower bound (up to constants) is non-adaptive, indicating that adaptivity does not help much in the detection problem.
- Finally, we also note that the procedure does not require measurement errors to be Gaussian. Via an application of Chebyshev's inequality, it suffices if the errors have finite second moment. We discuss this aspect further in Section 6.

3.2. Support recovery from passive measurements

In this section, we address the problem of estimating the support set B^* of the signal Θ^* from noisy linear measurements of the form specified in Equation (2.1). We provide a minimax lower bound on the support recovery risk for the case when the measurement matrices $(X_i)_{i\in[n]}$ are independent with jointly Gaussian entries drawn from a mean zero, spherical Gaussian, with variance $\frac{1}{d_1 d_2}$, i.e. each,

$$X_i \sim N\left(0, \frac{I}{d_1 d_2}\right) \qquad i \in [n]. \tag{3.2}$$

For our upper bounds we consider the case when the measurement matrices are drawn from a Σ-Gaussian ensemble. Here for a fixed $\Sigma \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$, each measurement matrix is distributed according to,

$$X_i \sim N\left(0, \frac{\Sigma}{\mathrm{tr}(\Sigma)}\right) \qquad i \in [n]. \tag{3.3}$$

In both cases the normalization ensures that $\mathbb{E}\|X_i\|_{\text{fro}}^2 = 1$. In order to capture the difficulty of support recovery from correlated measurements we consider the following quantity introduced in the work of Wainwright [45]:

$$\rho_{B^*}(\Sigma) = \min_{B \neq B^*, B \in \mathcal{B}} \lambda_{\min}\left(\Sigma_{(B^* \setminus B)B^* \setminus B} - \Sigma_{(B^* \setminus B)B}(\Sigma_{BB})^{-1}\Sigma_{B(B^* \setminus B)}\right), \quad (3.4)$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of its argument. In essence, the quantity $\rho_{B^*}(\Sigma)$ captures the maximal correlation between the measurement ensemble on the unknown true block B^* and any other block $B \in \mathcal{B}$. For the case of the standard Gaussian ensemble in Equation (3.2) we have that $\rho_{B^*}(\Sigma) = \rho_{B^*}(I) = 1$.

For our upper bounds we study the performance of a simple least-squares decoder. For a given block B, we denote by Θ_B a matrix in $\mathbb{R}^{d_1 \times d_2}$ whose support is restricted to be contained within B. With this definition in place and recalling the definition of $\mathcal{B}$ in Equation (2.2), we consider the following support estimator:

$$\widehat{B} = \arg\min_{B \in \mathcal{B}} \min_{\Theta_B} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle\!\langle \Theta_B, X_i \rangle\!\rangle)^2. \quad (3.5)$$

Computing this estimator requires solving roughly $d_1 d_2$ least squares programs over $k_1 k_2$ variables. With these preliminaries in place we have the following result:

Theorem 2. *Fix any* $0 < \alpha \leq 1/4$.

1. *Consider sensing matrices* X_i *drawn from the ensemble in Equation* (3.2). *There exists a positive universal positive constant* C *such that if,*

$$\frac{\theta_{\min}}{\sigma} \leq C(1 - 2\alpha)\sqrt{\frac{d_1 d_2}{n} \max\left(\frac{1}{\min(k_1, k_2)}, \frac{\log(d_1 d_2)}{k_1 k_2}\right)},$$

   then the minimax support recovery risk $R^{\text{sup}} \geq \alpha > 0$.
2. *Conversely, for sensing matrices drawn from the ensemble in Equation* (3.3) *there exist universal positive constants* $C_1, C_2 > 0$ *such that, if* $n \geq C_1 \max\{k_1 k_2 + \rho_{B^*}(\Sigma)^{-1}\theta_{\min}^{-2}\sigma^2, \log\max(d_1, d_2)\}$ *and*

$$\frac{\theta_{\min}}{\sigma} \geq C_2\sqrt{\frac{\text{tr}(\Sigma)\log(2/\alpha)}{n\rho_{B^*}(\Sigma)} \max\left(\frac{\log\max(k_1, k_2)}{\min(k_1, k_2)}, \frac{\log\max(d_1, d_2)}{k_1 k_2}\right)},$$

   then for the estimator in Equation (3.5) *we have that* $R^{\text{sup}}(\widehat{B}) \leq \alpha$.

Remarks:

- We can verify that in the case when $\Sigma = I$ we have that $\text{tr}(\Sigma) = d_1 d_2$, and that $\rho_{B^*}(\Sigma) = 1$ so that the upper and lower bounds match upto a $\log\max(k_1, k_2)$ factor in the worst case. More generally, the upper and lower bounds match upto a logarithmic factor for sufficiently well-conditioned designs, as captured by $\rho_{B^*}(\Sigma)$.

Algorithm 1 Approximate support recovery

input Measurement budget $n \geq \log\left(\frac{d_1 d_2}{4k_1 k_2}\right)$, a collection of size $d_1 d_2/(4k_1 k_2)$ of blocks $\mathcal{D}$ each of size $2k_1 \times 2k_2$.

Initial support: $J_0^{(1)} := \{1, \ldots, \frac{d_1 d_2}{4k_1 k_2}\}$, $s_0 := \log \frac{d_1 d_2}{4k_1 k_2}$.

For each s in $1, \ldots, \log_2 \frac{d_1 d_2}{4k_1 k_2}$

1. Allocate: $n_s := \lfloor (n - s_0) s 2^{-s-1} \rfloor + 1$.
2. Split: Partition $J_0^{(s)}$ into two collections of blocks of equal size, $J_1^{(s)}$ contains the first $\frac{d_1 d_2}{2^{s+2} k_1 k_2}$ blocks and $J_2^{(s)}$ the remainder.
3. Sensing matrix: $X_s = \sqrt{\frac{2^{-(s_0 - s + 1)}}{4k_1 k_2}}$ on $J_1^{(s)}$, $X_s = -\sqrt{\frac{2^{-(s_0 - s + 1)}}{4k_1 k_2}}$ on $J_2^{(s)}$ and 0 otherwise.
4. Measure: $y_i^{(s)} = \langle\langle \Theta^*, X_s \rangle\rangle + \epsilon_i^{(s)}$ for $i \in [1, \ldots, n_s]$.
5. Update support: $J_0^{(s+1)} = J_1^{(s)}$ if $\sum_{i=1}^{n_s} y_i^{(s)} > 0$ and $J_0^{(s+1)} = J_2^{(s)}$ otherwise.

output The single block in $J_0^{(s_0+1)}$.

- The proof of this theorem is given in Section 4. The lower bound in this case follows from the construction of two packing sets, and an application of Fano's inequality. The two packing sets capture the difficulty of precisely localizing the signal and approximately localizing the signal. These difficulties are reflected in the two terms in the lower bound. The upper bound uses a technical result of Wainwright [45] in order to bound the probability that the least squares estimator picks a fixed incorrect block. This result is then combined with a counting argument to control the overall probability of error. We note that our results generalize in a straightforward manner to sub-Gaussian designs. A direct application of results for sparse vector localization (see Table 1) would lose a factor of $1/\sqrt{\min(k_1, k_2)}$, which we gain by refining prior arguments to exploit the contiguous block structure. In the next section we show that even more substantial gains are possible when we choose measurements adaptively.
- While the decoder we analyze assumes knowledge of k_1, k_2 one can also use a similar procedure to adapt to the unknown size of the activation block. In particular, one can perform exhaustive search procedure for all possible sizes of activation blocks. Small modifications to our proof can be used to show that this procedure adapts to the unknown block size while still achieving the same risk. We omit the details but note that similar modifications have been detailed in [9] for a related problem.
- Finally, we note that the least squares decoder achieves the same result when the signal has mixed sign provided we define $\theta_{\min}$ to be the smallest absolute value of Θ^* over B^*. This is contrary to the testing procedure considered previously, and we re-visit this issue in Section 6.

3.3. Support recovery from adaptive measurements

In this section, we consider the problem of identifying the support set B^*, using adaptive linear measurements. To be clear, in this setting each measurement matrix X_i may be a function of $(y_j, X_j)_{j\in[i-1]}$. To simplify the exposition, in this section we assume that d_1 is a multiple of $2k_1$ and that d_2 is a multiple of $2k_2$. We provide an upper bound by analyzing the procedures described in Algorithms 1 and 2:

- **Algorithm 1:** At a high-level, Algorithm 1 takes as input a collection of non-overlapping blocks of size $2k_1 \times 2k_2$ and collects compressive measurements of these blocks following a compressed binary search procedure [19]. The compressed binary search procedure essentially divides the collection of blocks recursively into halves and uses a small modification of the detection procedure described in Section 3.1 in order to decide which half to measure further (see Figure 1). As output, Algorithm 1 produces a single candidate block which we take further measurements of in Algorithm 2. In order to further simplify the analysis we repeat Algorithm 1 on four staggered collections of blocks, one of which is guaranteed to have a block that fully contains B^*. In more detail, we run Algorithm 1 on the four collections:

$$\begin{aligned}
\mathcal{D}_1 := \{&B_{1,1} := [1,\dots,2k_1]\times[1,\dots,2k_2],\\
&B_{1,2} := [2k_1+1,\dots,4k_1]\times[1,\dots,2k_2],\\
&\dots, B_{1,d_1d_2/4k_1k_2} := [d_1-2k_1,\dots,d_1]\times[d_2-2k_2,\dots,d_2]\}\\
\mathcal{D}_2 := \{&B_{2,1} := [k_1,\dots,3k_1]\times[k_2,\dots,3k_2],\\
&B_{2,2} := [3k_1+1,\dots,5k_1]\times[k_2,\dots,3k_2],\\
&\dots, B_{2,d_1d_2/4k_1k_2} := [d_1-k_1,\dots,d_1,1,\dots,k_1]\times\\
&[d_2-k_2,\dots,d_2,1,\dots,k_2]\}\\
\mathcal{D}_3 := \{&B_{3,1} := [k_1,\dots,3k_1]\times[1,\dots,2k_2],\\
&B_{3,2} := [3k_1+1,\dots,5k_1]\times[1,\dots,2k_2]\\
&\dots, B_{3,d_1d_2/4k_1k_2} := [d_1-k_1,\dots,d_1,1,\dots,k_1]\times[d_2-2k_2,\dots,d_2]\},
\end{aligned}$$

  and

$$\begin{aligned}
\mathcal{D}_4 := \{&B_{4,1} := [1,\dots,2k_1]\times[k_2,\dots,3k_2],\\
&B_{4,2} := [2k_1+1,\dots,4k_1]\times[k_2,\dots,3k_2]\\
&\dots, B_{4,d_1d_2/4k_1k_2} := [d_1-2k_1,\dots,d_1]\times[d_2-k_2,\dots,d_2,1,\dots,k_2]\}.
\end{aligned}$$

  Figure 2 illustrates the four collections. Effectively, upon completion of Algorithm 1 we have a collection of $8k_1$ row indices and $8k_2$ column indices that with high-probability contain B^*, i.e., we have approximately localized B^*.
- **Algorithm 2:** We use Algorithm 2 to precisely identify the k_1 row indices and k_2 column indices that compose B^*, from the $8k_1$ row indices

Algorithm 2 Exact recovery (of columns)

input Measurement budget n, a sub-matrix $B \in \mathbb{R}^{8k_1 \times 8k_2}$, success probability α. Set $\widetilde{n} = n/36$.

1. Measure selected columns: $y_i^c = (8k_1)^{-1/2} \sum_{l=1}^{8k_1} B_{lc} + \epsilon_i^c$, for $c = \{1, k_2 + 1, 2k_2 + 1, \ldots, 7k_2 + 1\}$, each $\widetilde{n}$ times.
2. Let $l = \operatorname{argmax}_c \sum_{i=1}^{\widetilde{n}} y_i^c$, $r = l + k_2$, $n_b = \lfloor \frac{\widetilde{n}}{3 \log_2 k_2} \rfloor$.
3. While $r - l \geq 1$
   (a) Let $c = \lfloor \frac{r+l}{2} \rfloor$.
   (b) Measure $y_i^c = (8k_1)^{-1/2} \sum_{l=1}^{8k_1} B_{lc} + \epsilon_i^c$ for $i = \{1, \ldots, n_b\}$.
   (c) If[a] $\sum_{i=1}^{n_b} y_i^c \geq \mathcal{O}\left(\sqrt{n_b \sigma^2 \log\left(\frac{\log k_2}{\alpha}\right)}\right)$ then $l = c$, otherwise $r = c$.

output Set of columns $\{l - k_2 + 1, \ldots, l\}$.

[a]The exact constants appear in the proof of Theorem 3.

and $8k_2$ column indices output by Algorithm 1. In the first stage of Algorithm 2 we measure repeatedly a small number of columns, exactly one of which is contained in B^*, in order to identify an active column with high probability. The next stage finds the first non-active column to the left and right by testing columns using a binary search procedure. In this way, all the active columns are located. Finally, Algorithm 2 is repeated on rows in order to completely localize B^*.

In summary, our support recovery estimator runs Algorithm 1 on the 4 collections described above, and then runs Algorithm 2 on the set of $8k_2$ indices returned by Algorithm 1 in order to identify k_2 columns. Algorithm 2 is then repeated mutatis mutandis to identify k_1 rows. The following result characterizes both the fundamental limits on any adaptive procedure as well as the performance of the estimator we have described:

Theorem 3. *Fix any* $0 < \alpha \leq 1/4$.

1. *There is a universal constant* $C_1 > 0$ *such that if*
$$\frac{\theta_{\min}}{\sigma} < C_1(1 - 2\alpha) \max\left(\sqrt{\frac{d_1 d_2}{n k_1^2 k_2^2}}, \sqrt{\frac{1}{n \min(k_1, k_2)}}\right)$$
then $R^{\sup} \geq \alpha$, *for any adaptive procedure using* n *measurements.*
2. *Conversely, there is a universal constant* $C_2 > 0$ *such that if* $n \geq \log(d_1 d_2)$ *and if*
$$\frac{\theta_{\min}}{\sigma} \geq C_2 \sqrt{\log \frac{1}{\alpha}} \left(\max\left(\sqrt{\frac{d_1 d_2}{n k_1^2 k_2^2}}, \sqrt{\frac{\log\left(\max(k_1, k_2)\right) \log\log(k_1 k_2)}{n \min(k_1, k_2)}}\right)\right)$$
then the estimator described above has risk $R^{\sup}(\widehat{B}) \leq \alpha$.

Remarks:

- Once again for our lower bound we construct specific pairs of hypotheses that are hard to distinguish. The two terms in the lower bound reflect the hardness of estimating the support of the signal in two important cases. The first term reflects the difficulty of approximately estimating B^*. This term grows at the same rate as the detection lower bound, and its proof is similar. Given a coarse estimate of the support, we still need to identify the exact support of the signal. The hardness of this task gives rise to the second term in the lower bound. This term is independent of d_1 and d_2, as is to be expected, but has a considerably worse dependence on k_1 and k_2. From a technical standpoint we note that since the lower bounds need to account for possibly adaptive strategies, we cannot apply Fano's inequality (see for instance Arias-Castro et al. [3] for a discussion) in a straightforward way. Instead we rely on direct arguments based on pairs of hypotheses as opposed to packing sets.
- The upper bound matches the lower bound up to a logarithmic $\mathcal{O}\left(\sqrt{\log\left(\max(k_1,k_2)\right)\log\log(k_1k_2)}\right)$ factor. It is worth noting that for small k_1 and k_2, when the first term of the upper bound dominates, our adaptive support recovery procedure achieves the detection limits (see Theorem 1). When the support is larger, the lower bound indicates that no procedure can achieve the detection rate.
- We can further compare the results of this theorem to the best possible result for passive procedures given in Theorem 2. Our adaptive procedure is significantly more efficient than any passive procedure, and is able to recover the support of signals that are weaker by a $\sqrt{d_1d_2}$ factor. Unlike in the weakly structured sparse vector case, the great potential for gains from adaptive measurements is clearly seen in the highly structured contiguous block setting. This in turn highlights the fundamental interplay between structure and adaptivity.

4. Proofs of main results

In this section, we prove our three main theorems, while deferring more technical aspects of the proofs to the Appendix. Throughout the proofs, we will denote by $c_1, c_2, \ldots$ positive constants that may change their value from line to line.

4.1. Proof of Theorem 1

Proof of the lower bound: We first reduce the space of alternatives to only include matrices whose entries are exactly $\theta_{\min}$, i.e., for a given value $\theta_{\min} > 0$ we define the class of matrices:

$$\widetilde{\mathbf{\Theta}}(\theta_{\min}) = \{\Theta : \exists\ B \in \mathcal{B},\ \text{such that supp}(\Theta) = B, \Theta_{ij} = \theta_{\min}\ \forall\ (i,j) \in B\}\,,$$

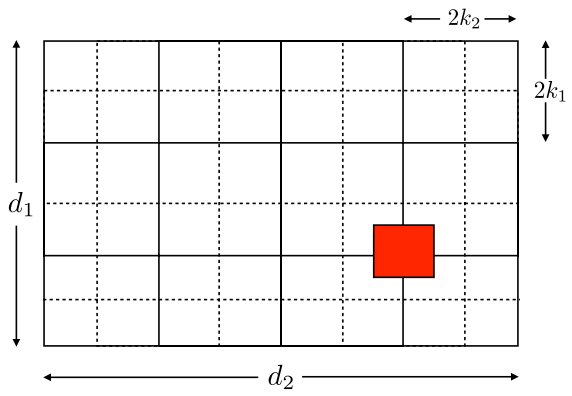

FIG 2. *The collection of blocks $\mathcal{D}_1$ is shown in solid lines and the collection $\mathcal{D}_2$ is shown in dashed lines. The collections $\mathcal{D}_3$ and $\mathcal{D}_4$ overlap with these and are not illustrated in the figure. The $(k_1 \times k_2)$ block of activation B^* is shown in red.*

and consider the hypothesis testing problem:

$$\begin{aligned} H_0 &: \quad \Theta^* = 0_{d_1 \times d_2} \\ H_1 &: \quad \Theta^* \in \widetilde{\boldsymbol{\Theta}}(\theta_{\min}). \end{aligned}$$

This hypothesis testing problem does not a have a simple alternative. We use a classical reduction to a simple versus simple hypothesis testing problem by considering a prior π over the alternatives, and noting that the Bayes detection risk under any prior π lower bounds the minimax detection risk. In our setting a uniform prior over the alternatives will suffice. Concretely, letting π denote the uniform measure over $\widetilde{\boldsymbol{\Theta}}(\theta_{\min})$ we focus on the following hypothesis testing problem:

$$\begin{aligned} H_0 &: \quad \Theta^* = 0_{d_1 \times d_2} \\ H_1 &: \quad \Theta^* \sim \pi. \end{aligned}$$

Appealing to the classical Neyman-Pearson lemma, it suffices to consider the optimal likelihood-ratio test and its risk is lower bounded via a standard argument (see for instance Arias-Castro [2]) as:

$$R^{\text{det}} \geq 1 - \frac{1}{2}\sqrt{\frac{\text{KL}(\mathbb{P}_0((y_i, X_i)_{i\in[n]}), \mathbb{E}_{\Theta\sim\pi}\mathbb{P}_\Theta((y_i, X_i)_{i\in[n]}))}{2}},$$

where KL denotes the KL divergence. It thus remains to upper bound the KL divergence. We note that care is needed in upper bounding the KL divergence since our bound needs to apply to both passive and adaptive sensing schemes and thus for instance standard tensorization arguments are invalid. We have the following technical lemma:

Lemma 1. *Under the uniform measure π over $\widetilde{\mathbf{\Theta}}(\theta_{\min})$ for any possibly adaptive sensing scheme we have that,*

$$\mathrm{KL}(\mathbb{P}_0((y_i, X_i)_{i\in[n]}), \mathbb{E}_{\Theta\sim\pi}\mathbb{P}_\Theta((y_i, X_i)_{i\in[n]})) \leq \frac{n\theta_{\min}^2 k_1^2 k_2^2}{8\sigma^2 d_1 d_2}.$$

We prove this lemma in the Appendix. Taking this claim as given, we can now complete the proof of the theorem. Some simple algebra shows that,

$$R^{\mathrm{det}} \geq 1 - \frac{\theta_{\min} k_1 k_2}{\sigma}\sqrt{\frac{n}{64 d_1 d_2}},$$

which in turn is at least α when,

$$\frac{\theta_{\min}}{\sigma} \leq (1-\alpha)\sqrt{\frac{64 d_1 d_2}{n k_1^2 k_2^2}},$$

as claimed. Finally, we conclude the proof of the theorem by noting that the upper bound follows directly from Proposition 1 of Arias-Castro [2].

4.2. Proof of Theorem 2

Proof of the lower bound: Our proof proceeds via the construction of two packing sets. At a high level, the first construction is intended to capture the difficulty of distinguishing between the true block and blocks that that overlap with the true block on all but one row/column. The second construction deals with distinguishing between the true block and the large collection of blocks that do not overlap with the first one. Without loss of generality we assume $k_1 \leq k_2$.

Construction 1: We consider two distributions $\mathbb{P}_{\Theta_1}$ and $\mathbb{P}_{\Theta_2}$, where $B_1 = \mathrm{supp}(\Theta_1) = [1,\ldots,k_1]\times[1,\ldots,k_2]$ and $B_2 = \mathrm{supp}(\Theta_2) = [1,\ldots,k_1]\times[2,\ldots,k_2+1]$ and every non-zero element of Θ_1 and Θ_2 are equal to $\theta_{\min}$. Observe that the two supports differ in only one column. As in the detection problem we use a classical reduction to testing, and lower bound the testing error via the KL divergence:

$$\begin{aligned} R^{\mathrm{sup}}(\Psi) &= \sup_{\Theta\in\mathbf{\Theta}(\theta_{\min})} P_\Theta\left[\Psi\big((y_i, X_i)_{i\in[n]}\big) \neq \mathrm{supp}(\Theta)\right], \\ &\geq \max_{j\in\{1,2\}} \mathbb{P}_{\Theta_j}\left[\Psi\left((y_i, X_i)_{i\in[n]}\right) \neq B_j\right] \\ &\geq \frac{1}{2}\left(1 - \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\Theta_1}, \mathbb{P}_{\Theta_2})}{8}}\right). \end{aligned}$$

Therefore, we need to compute an upper bound on the KL divergence. The first

bound in the theorem follows from the following calculation

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\Theta_1}, \mathbb{P}_{\Theta_2}) &= \mathbb{E}_{\mathbb{P}_{\Theta_1}} \log \frac{\mathbb{P}_{\Theta_1}}{\mathbb{P}_{\Theta_2}} \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_{\Theta_1}} \sum_{i=1}^{n} \left(\langle\!\langle \Theta_2,\ X_i \rangle\!\rangle - \langle\!\langle \Theta_1,\ X_i \rangle\!\rangle \right)^2 \\ &= \frac{n}{\sigma^2} \frac{\|\Theta_1 - \Theta_2\|_{\mathrm{fro}}^2}{d_1 d_2} \\ &= \frac{n\theta_{\min}^2 k_1}{\sigma^2 d_1 d_2},\end{aligned} \tag{4.1}$$

where we have used the fact that X_i is a Gaussian random matrix with independent entries of variance $\frac{1}{d_1 d_2}$. Using this upper bound on the KL divergence in the expression above leads to the first term in the lower bound.

Construction 2: We consider a collection of distributions $\mathbb{P}_{\Theta_1}, \ldots, \mathbb{P}_{\Theta_{t+1}}$ where $t = (d_1 - k_1)(d_2 - k_2)$. The distribution $\mathbb{P}_{\Theta_1}$ is the same as in the first construction, while supports of signals $\Theta_2, \ldots, \Theta_{t+1}$ do not overlap with the support of the signal Θ_1. In order to obtain a lower bound on the risk, we use Fano's inequality (see, for example, Theorem 2.5 in [42]). In order to apply the inequality, we need to upper bound the KL divergence between $\mathbb{P}_{\Theta_1}$ and every $\mathbb{P}_{\Theta_j}$. As in the case of Construction 1, we now obtain:

$$\mathrm{KL}(\mathbb{P}_{\Theta_1}, \mathbb{P}_{\Theta_j}) \leq \frac{\theta_{\min}^2}{\sigma^2} \frac{n k_1 k_2}{d_1 d_2} \qquad \forall\ \ j \in \{2, \ldots, t+1\}. \tag{4.2}$$

Via an application of Fano's inequality we have that if:

$$\frac{\frac{\theta_{\min}^2}{\sigma^2} \frac{n k_1 k_2}{d_1 d_2} + \log 2}{\log((d_1 - k_1)(d_2 - k_2))} \leq (1 - 2\alpha),$$

then the support recovery risk is at least α. This leads to the second term in the lower bound of the theorem.

Proof of the upper bound: At a high-level we follow the technique of Wainwright [45] to analyze the performance of the least squares decoder. This involves first establishing an upper bound on the probability of incorrectly selecting a certain block $B \neq B^*$ in terms of the size of their non-overlap $|B^* \backslash B|$, and then combining these via a counting argument.

For a given block B we denote by $X_B \in \mathbb{R}^{n \times |B|}$, the measurement matrices restricted to the block B. We can then define the projection operator:

$$\Pi_B = X_B (X_B^T X_B)^{-1} X_B^T,$$

and the excess error of a subset B relative to the true subset B^*:

$$\Delta(B; B^*) = \|(I - \Pi_B) y\|_2^2 - \|(I - \Pi_{B^*}) y\|_2^2.$$

It is then straightforward to see that the least squares decoder fails if $\Delta(B; B^*) < 0$ for any $B \neq B^*$. For a block B we further define a measure of its overlap with B^* that takes into account the correlations in the measurements. Recalling, the definition of $\rho_{B^*}(\Sigma)$ in Equation (3.4) we define:

$$\Phi(B) := \frac{\rho_{B^*}(\Sigma)|B^* \backslash B|\theta_{\min}^2}{\mathrm{tr}(\Sigma)\sigma^2}.$$

Finally, throughout the rest of this proof we denote $\widetilde{n} = n - k_1 k_2$. With these definitions in place we now state a technical lemma from Wainwright [45] (see Lemma 2):

Lemma 2. *There is a universal constant C such that as long as $\widetilde{n} \geq \rho_{B^*}(\Sigma)^{-1}\theta_{\min}^{-2}\sigma^2$, for any fixed block B we have that:*

$$\mathbb{P}(\Delta(B; B^*) < 0) \leq 4 \exp\left(-\frac{\widetilde{n}\Phi(B)}{64(\Phi(B) + 8)}\right).$$

With this lemma in place it remains to analyze the error probability of the least squares decoder. Due to the contiguous block structure of our signal we do this analysis differently from previous work and this leads to sharp results in our setting.

Our first step is to count the number of blocks in $\mathcal{B}$ which have a specified non-overlap with B^*. We denote by $N(t)$ the number of blocks B that have non-overlap $|B^* \backslash B| = t$. In order to upper bound the failure probability of the least squares decoder we combine the above calculation and Lemma 2 via the union bound. Concretely, we can write the probability of error as:

$$\begin{aligned} p_e &:= \mathbb{P}(\exists\ B \in \mathcal{B}, B \neq B^* \text{ such that } \Delta(B; B^*) < 0) \\ &\leq 4 \sum_{B \in \mathcal{B}, B \neq B^*} \exp\left(-\frac{\widetilde{n}\Phi(B)}{64(\Phi(B) + 8)}\right). \end{aligned}$$

This is turn can be written as:

$$p_e \leq 4 \sum_t N(t) \exp\left(-\frac{\widetilde{n} t \rho_{B^*}(\Sigma)\theta_{\min}^2}{64(t\rho_{B^*}(\Sigma)\theta_{\min}^2 + 8\mathrm{tr}(\Sigma)\sigma^2)}\right).$$

Observe that due to the highly structured nature of our class of signals $N(t)$ can only take a limited number of values, particularly it is easy to see that

$$N(k_1 k_2) = (d_1 - k_1)(d_2 - k_2).$$

Furthermore for each value of the form (i, j) with $i \in \{1, \ldots, k_1\}$ and $j \in \{1, \ldots, k_2\}$ with $(i, j) \neq (1, 1)$ we have that

$$N(k_1 k_2 - (k_1 - i + 1)(k_2 - j + 1)) \leq 4.$$

Using this we have that,

$$p_e \leq 16 \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \mathbb{I}((i,j) \neq (1,1)) \times \\ \exp\left(-\frac{\widetilde{n}(k_1k_2 - (k_1 - i + 1)(k_2 - j + 1))\rho_{B^*}(\Sigma)\theta^2_{\min}}{64((k_1k_2 - (k_1 - i + 1)(k_2 - j + 1))\rho_{B^*}(\Sigma)\theta^2_{\min} + 8\mathrm{tr}(\Sigma)\sigma^2)} \right) \\ +4(d_1 - k_1)(d_2 - k_2) \exp\left(-\frac{\widetilde{n}k_1k_2\rho_{B^*}(\Sigma)\theta^2_{\min}}{64(k_1k_2\rho_{B^*}(\Sigma)\theta^2_{\min} + 8\mathrm{tr}(\Sigma)\sigma^2)} \right).$$

We can treat these two terms separately. First, observe that if

$$\frac{\theta_{\min}}{\sigma} \geq C \sqrt{\frac{\mathrm{tr}(\Sigma)\log(1/\alpha)}{\rho_{B^*}(\Sigma)\widetilde{n}} \frac{\log(d_1d_2)}{k_1k_2}},$$

then under the conditions on the sample-size we have that the second term $< \alpha/2$. It remains to analyze the first term. Without loss of generality we can assume that $k_1 \leq k_2$. By upper bounding each term in the sum with the largest term we have that,

$$\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \mathbb{I}((i,j) \neq (1,1)) \times \\ \exp\left(-\widetilde{n}\frac{(k_1k_2 - (k_1 - i + 1)(k_2 - j + 1))\rho_{B^*}(\Sigma)\theta^2_{\min}}{64((k_1k_2 - (k_1 - i + 1)(k_2 - j + 1))\rho_{B^*}(\Sigma)\theta^2_{\min} + 8\mathrm{tr}(\Sigma)\sigma^2)} \right) \\ \leq k_2^2 \exp\left(-\frac{\widetilde{n}k_1\rho_{B^*}(\Sigma)\theta^2_{\min}}{64(k_1\rho_{B^*}(\Sigma)\theta^2_{\min} + 8\mathrm{tr}(\Sigma)\sigma^2)} \right),$$

which is smaller than $\alpha/2$ when

$$\frac{\theta_{\min}}{\sigma} \geq C \sqrt{\frac{\mathrm{tr}(\Sigma)\log(1/\alpha)}{\rho_{B^*}(\Sigma)\widetilde{n}} \frac{\log(\max(k_1,k_2))}{\min(k_1,k_2)}},$$

and this completes the proof.

4.3. Proof of Theorem 3

Proof of the lower bound: The proof will proceed via two separate constructions. Motivated by similar considerations as in the passive measurements setting we use different constructions to capture the difficulty of approximately and exactly localizing the support of the signal. However in this setting further care is needed to deal with adaptive measurement schemes.

Construction 1: We assume that d_1 and d_2 are multiples of two to simplify the exposition. We first define two sets:

$$\mathcal{B}_1 = \left\{ I_r \times I_c \ : \ \begin{array}{l} I_r \text{ and } I_c \text{ are contiguous subsets of } [d_1/2] \text{ and } [d_2/2], \\ |I_r| = k_1, |I_c| = k_2 \end{array} \right\},$$

$$\mathcal{B}_2 = \left\{ I_r \times I_c \ : \ \begin{array}{c} I_r \text{ and } I_c \text{ are contiguous subsets of } \{d_1/2, \ldots, d_1\} \\ \text{and } \{d_2/2, \ldots, d_2\}, |I_r| = k_1, |I_c| = k_2 \end{array} \right\}.$$

Now, we define two distributions which correspond to the cases when $B^* \in \mathcal{B}_1$ and when $B^* \in \mathcal{B}_2$ respectively. Let π_1 and π_2 denote the uniform measure on $\mathcal{B}_1$ and $\mathcal{B}_2$ respectively. Then, $\mathbb{P}_1 = \mathbb{E}_{\Theta \sim \pi_1} \mathbb{P}_\Theta$ and $\mathbb{P}_2 = \mathbb{E}_{\Theta \sim \pi_2} \mathbb{P}_\Theta$. The following lemma proved in the Appendix upper bounds the KL divergence between the two distributions (allowing for adaptive sensing schemes):

Lemma 3. *For any possibly adaptive sensing scheme we have that,*

$$\mathrm{KL}(\mathbb{P}_1((y_i, X_i)_{i \in [n]}), \mathbb{P}_2((y_i, X_i)_{i \in [n]})) \leq \frac{n\theta_{\min}^2 k_1^2 k_2^2}{4\sigma^2 d_1 d_2}.$$

With this lemma in place we obtain the desired lower bound by noting that the minimax risk $R^{\sup}$ for distinguishing $\mathbb{P}_1$ from $\mathbb{P}_2$ is lower bounded as:

$$R^{\sup} \geq \frac{1}{2}\left(1 - \frac{1}{2}\sqrt{\frac{\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2)}{2}}\right).$$

From this we obtain that if

$$\frac{\theta_{\min}}{\sigma} \leq c_1(1-2\alpha)\sqrt{\frac{d_1 d_2}{n k_1^2 k_2^2}},$$

then $R^{\sup} \geq \alpha$ as desired.

Construction 2: Consider, two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, where $\mathbb{P}_1$ is induced by matrix Θ_1 with support $B = B_1 = [1, \ldots, k_1][1, \ldots, k_2]$, and $\mathbb{P}_2$ is induced by matrix Θ_2 with support $B = B_2 = [1, \ldots, k_1][2, \ldots, k_2 + 1]$. We set matrices Θ_1 and Θ_2 to have non-zero elements all equal to $\theta_{\min}$.

Once again we need to upper bound the KL divergence between these two distributions while allowing for adaptive sensing schemes. We prove the following lemma in the Appendix:

Lemma 4. *For any possibly adaptive sensing scheme we have that,*

$$\mathrm{KL}(\mathbb{P}_1((y_i, X_i)_{i \in [n]}), \mathbb{P}_2((y_i, X_i)_{i \in [n]})) \leq \frac{n\theta_{\min}^2 \min(k_1, k_2)}{4\sigma^2 d_1 d_2}.$$

Once again we can use our earlier lower bound on the minimax risk together with this lemma to obtain:

$$R^{\sup} \geq \frac{1}{2}\left(1 - \sqrt{\frac{\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2)}{8}}\right) \geq \frac{1}{2}\left(1 - \sqrt{\frac{n \min(k_1, k_2)\theta_{\min}^2}{8\sigma^2}}\right).$$

This gives that if

$$\frac{\theta_{\min}}{\sigma} \leq c_2(1-2\alpha)\sqrt{\frac{1}{\min(k_1, k_2)}},$$

then $R^{\sup} \geq \alpha$. The proof of the lower bound is completed by combining the lower bounds from the two constructions.

Proof of the upper bound: As a preliminary, we note a standard Gaussian tail bound that we will use repeatedly in our proof: Let $Z \sim \mathcal{N}(0,1)$ be a standard Normal random variable. Then for $t > 0$

$$\mathbb{P}(Z > t) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp(-t^2/2). \tag{4.3}$$

In the sequel we analyze the performance of the two algorithms that constitute our support estimator. To ease presentation we will assume d_1 is a dyadic multiple of $2k_1$ and d_2 a dyadic multiple of $2k_2$. Straightforward modifications are possible when this is not the case. Further, we suppose that the measurement budget is divided equally between the two Algorithms, i.e., each of Algorithm 1 and 2 use $n/2$ samples.

Algorithm 1 is run on the four collections of blocks introduced in Section 3.3. Of these four collections, at least one collection contains a block that completely contains B^*. Algorithm 1 succeeds if when run on that particular collection it returns the block that completely contains B^*, irrespective of its output on the remaining three collections. The following lemma gives sufficient conditions for the success of Algorithm 1:

Lemma 5. *For any $0 < \alpha < 1$, Algorithm 1 succeeds with probability at least $1 - \alpha$ if,*

$$\frac{\theta_{\min}}{\sigma} \geq \sqrt{\frac{128 d_1 d_2}{n_1 k_1^2 k_2^2} \log\left(\frac{16}{\alpha}\right)}.$$

We prove this lemma in the Appendix by analyzing the performance of the binary search procedure used by Algorithm 1. In the analysis of Algorithm 2, we condition on the success of Algorithm 1, and note in passing that there are no dependence issues since each Algorithm collects its own measurements. Algorithm 2 is run twice, once on the collection of $8k_1$ row indices and once on the collection of $8k_2$ column indices returned by Algorithm 1. We say that Algorithm 2 succeeds when it outputs the correct support B^*.

Lemma 6. *For any $0 < \alpha < 1$, Algorithm 2 succeeds with probability at least $1 - \alpha$ if,*

$$\frac{\theta_{\min}}{\sigma} \geq \sqrt{\frac{1152 \log \max(k_1, k_2)}{n \min(k_1, k_2)} \log\left(\frac{24 \log \max(k_1, k_2)}{\alpha}\right)}.$$

We give a detailed proof of this lemma in the Appendix. The result follows from a careful analysis of the two phases of Algorithm 2. The first phase measures a selected set of rows/columns repeatedly in order to find a single row/column contained in B^* while the second phase then uses binary search in order to locate the first and last indices of the row/column support. The theorem follows by combining Lemma 5 and 6 via the union bound.

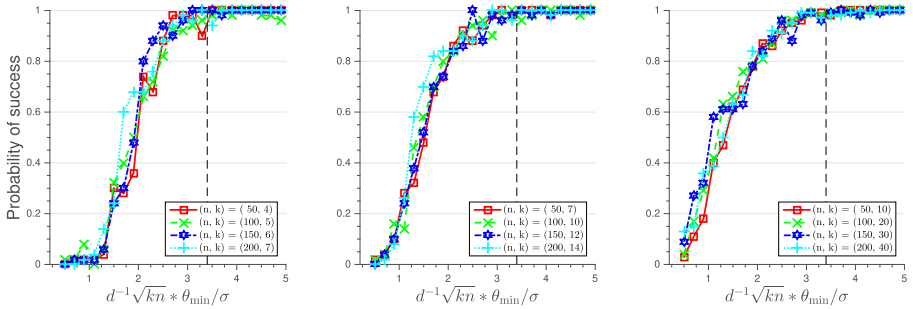

FIG 3. *Probability of success with passive measurements (averaged over 100 simulation runs).*

5. Simulations

In this section we provide some simulation results to illustrate the finite sample performance of our proposed procedures. Concretely, Theorem 2 and Theorem 3 characterize the signal amplitude needed for the passive and adaptive identification of a contiguous block, respectively. We demonstrate that the scalings predicted by these theorems are sharp by plotting the probability of successful recovery against an appropriately rescaled signal amplitude and showing that, as predicted, these curves for line up for different values of the problem parameters. For simplicity, we let $d_1 = d_2 = d$ and $k_1 = k_2 = k$.

5.1. Support recovery from passive measurements

In our first set of simulations, we plot the probability of success of the least squares decoder against $d^{-1}\sqrt{kn} * (\theta_{\min}/\sigma)$, where the number of measurements $n = 100$, averaged over 100 simulation runs. Each plot in Figure 3 represents different relationship between k and d; in the first plot, $k = \Theta(\log d)$, in the second $k = \Theta(\sqrt{d})$, while in the third plot $k = \Theta(d)$. The dashed vertical line denotes the scaled signal-to-noise ratio at which the probability of success is larger than 0.95. We observe that irrespective of the problem size and the relationship between d and k, Theorem 2 accurately characterizes the minimum signal amplitude needed for successful identification.

5.2. Support recovery from adaptive measurements

In our second set of simulations, we consider the probability of successful localization of B^* using the procedure outlined in Section 3.3, and $n = 500$ adaptively chosen measurements. The signal-to-noise ratio $(\theta_{\min}/\sigma)$ is scaled by $d^{-1}\sqrt{n}k^2$ in the first two plots where $k = \Theta(\log d)$ and $k = \Theta(\sqrt{d})$ respectively, while in the third plot the amplitude is scaled by $\sqrt{nk/\log k}$ since $k = \Theta(d)$. In the first two regimes, the cost of approximate localization dominates the rate, while in

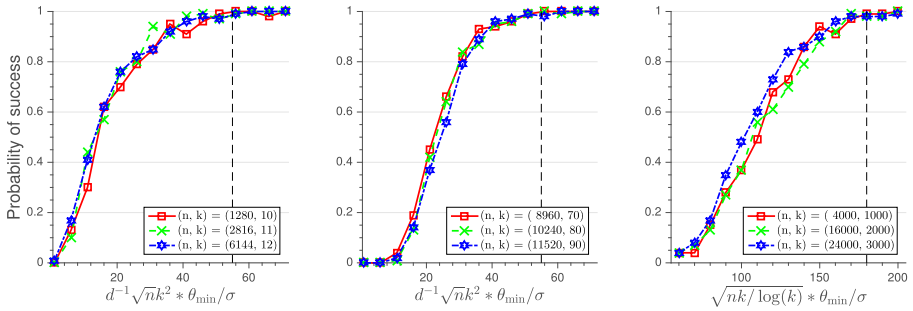

FIG 4. *Probability of success with adaptive measurements (averaged over 100 simulation runs).*

the third regime the cost of exact localization dominates the rate. Once again we observe that Theorem 3 sharply characterizes the minimum signal amplitude needed for successful identification.

6. Extensions and discussion

In this paper, we establish the fundamental limits for the problem of detecting and localizing a contiguous block of weak activation in a data matrix from either adaptive or non-adaptive compressive measurements. Our bounds characterize the tradeoffs between the signal amplitude, size of the matrix, size of the non-zero block and number of measurements. We also demonstrate computationally efficient estimators that achieve these bounds.

There are several extensions of the work presented in this paper that would be interesting to consider in future work. It should be possible to extend our analysis from the preceding sections so as to obtain similar results for the case when the size of B^* is unknown. The problem of designing adaptive estimators in structured passive settings have been considered for instance by Butucea and Ingster [9]. Throughout this paper we focussed on the setting where $\theta_{\min} > 0$, i.e., on positive signals. In the passive setting, the least squares estimator has the same rate for mixed sign signals. In the adaptive setting one can use techniques from the work of Arias-Castro [2] to deal with mixed sign signals, albeit at the cost of an unavoidable $\sqrt{k_1 k_2}$ factor. While our paper focused on exact support recovery it would also be interesting to consider support recovery for instance in the Hamming metric. An examination of our proofs for detection and for support recovery from adaptive measurements reveals that these algorithms rely on the sub-Gaussian behaviour of simple averages of the matrix Θ^*. These algorithms can be modified to use robust estimates of these quantities (see for instance [17]) in order to deal with possibly heavy-tailed noise.

Finally, the contiguous block model we consider is one useful form of the more general class of structured sparse signals, where adaptivity helps considerably. Techniques developed in this work have since been applied to the problems of adaptive compressive sensing of other structured signals in the works of Soni

and Haupt [39], Krishnamurthy et al. [31] and Castro and Tánczos [16]. We expect a full characterization of structured signals for which adaptive sensing can be useful will be a fruitful path for future investigation.

Appendix A: Additional technical results

In this supplementary section, we provide proofs of the remaining technical claims.

A.1. Proof of Lemma 1

We define the likelihood ratio as:

$$L \equiv \frac{\mathbb{E}_{\Theta\sim\pi}\mathbb{P}_\Theta[(y_i, X_i)_{i\in[n]}]}{\mathbb{P}_0[(y_i, X_i)_{i\in[n]}]} = \frac{\mathbb{E}_{\Theta\sim\pi}\prod_{i=1}^n \mathbb{P}_\Theta[y_i|X_i]}{\prod_{i=1}^n \mathbb{P}_0[y_i|X_i]},$$

where the second equality follows by decomposing the probabilities by the chain rule and observing that $\mathbb{P}_0[X_i|(y_j, X_j)_{j\in[i-1]}] = \mathbb{P}_\Theta[X_i|(y_j, X_j)_{j\in[i-1]}]$, since the sampling strategy (whether active or passive) cannot depend on the (unknown) true hypothesis. This in turn can be written as:

$$L = \mathbb{E}_{\Theta\sim\pi}\exp\left(\sum_{i=1}^m \frac{2y_i\langle\!\langle \Theta,\ X_i\rangle\!\rangle - \langle\!\langle \Theta,\ X_i\rangle\!\rangle^2}{2\sigma^2}\right).$$

The likelihood ratio is in turn related to the KL divergence as:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0, \mathbb{E}_{\Theta\sim\pi}\mathbb{P}_\Theta) &= -\mathbb{E}_0 \log L \\
&\le -\mathbb{E}_{\Theta\sim\pi}\sum_{i=1}^n \mathbb{E}_0 \frac{2y_i\langle\!\langle \Theta,\ X_i\rangle\!\rangle - \langle\!\langle \Theta,\ X_i\rangle\!\rangle^2}{2\sigma^2} \\
&= \mathbb{E}_{\Theta\sim\pi}\sum_{i=1}^n \mathbb{E}_0 \frac{\langle\!\langle \Theta,\ X_i\rangle\!\rangle^2}{2\sigma^2} \\
&\le \frac{n}{2\sigma^2}\sup_{\|X\|_{\mathrm{fro}}\le 1} \mathbb{E}_{\Theta\sim\pi}\langle\!\langle \Theta,\ X\rangle\!\rangle^2 \\
&\le \frac{n}{2\sigma^2}\|\mathbb{E}_{\Theta\sim\pi}\mathrm{vec}(\Theta)\mathrm{vec}(\Theta)^T\|_{\mathrm{op}},
\end{aligned}$$

where the first inequality follows by applying the Jensen's inequality followed by Fubini's theorem, the intermediate equality follows by observing that under the null $y_i = \epsilon_i$ irrespective of X_i and the second inequality follows using the fact that $\|X_i\|_{\mathrm{fro}} = 1$. Denote by $C \in \mathbb{R}^{d_1d_2\times d_1d_2}$ the matrix:

$$C := \mathbb{E}_{\Theta\sim\pi}\mathrm{vec}(\Theta)\mathrm{vec}(\Theta)^T.$$

It remains only to bound the operator norm of C. The matrix C has diagonal elements $C_{ii} = \theta_{\min}^2\mathbb{E}_{\Theta\sim\pi}P_\Theta[\Theta_{\tau(i)} = 1]$ and off-diagonal elements $C_{ij} =$

$\theta_{\min}^2 \mathbb{E}_{\Theta\sim\pi} P_\Theta[\Theta_{\tau(i)} = 1, \Theta_{\tau(j)} = 1]$, where τ is an invertible map from a linear index in $\{1, \ldots, d_1 d_2\}$ to an entry of Θ. A bound on the operator norm of C follows from the following two observations. The support of the signal forms a contiguous block and therefore in any row of C there are at most $k_1 k_2$ non-zero entries. Furthermore, each non-zero entry in C is of magnitude at most $\theta_{\min}^2 k_1 k_2/(d_1 - k_1)(d_2 - k_2)$. Combining these two observations, we have

$$\|\!|C|\!\|_{\text{op}} \leq \max_j \sum_k |C_{jk}| \leq \frac{\theta_{\min}^2 k_1^2 k_2^2}{(d_1 - k_1)(d_2 - k_2)} \leq \frac{\theta_{\min}^2 k_1^2 k_2^2}{4 d_1 d_2}$$

where the last inequality follows from the assumption we maintain throughout the paper that $k_1 \leq d_1/2$ and $k_2 \leq d_2/2$. This completes the proof of the lemma.

A.2. Proof of Lemma 3

We first introduce another distribution $\mathbb{P}_0$ corresponding to the case when we have no signal ($\Theta^* = 0$). Now notice that using the triangle inequality for TV and Pinsker's inequality we obtain:

$$\begin{aligned}\|\mathbb{P}_1 - \mathbb{P}_2\|_{\text{TV}}^2 &\leq 2\|\mathbb{P}_0 - \mathbb{P}_1\|_{\text{TV}}^2 + 2\|\mathbb{P}_0 - \mathbb{P}_2\|_{\text{TV}}^2 \\ &\leq \text{KL}(\mathbb{P}_0, \mathbb{P}_1) + \text{KL}(\mathbb{P}_0, \mathbb{P}_2).\end{aligned}$$

Notice that $\text{KL}(\mathbb{P}_0, \mathbb{P}_1)$ is exactly the quantity we would have to upper bound to produce a lower bound on the signal strength for detecting whether a support of the signal is in the left half of the matrix or not (see Theorem 1). We can now apply a slight modification of the proof of Lemma 1 to obtain that

$$\text{KL}(\mathbb{P}_0, \mathbb{P}_1) = \text{KL}(\mathbb{P}_0, \mathbb{P}_2) \leq \frac{n\theta_{\min}^2 k_1^2 k_2^2}{8\sigma^2 d_1 d_2},$$

which in turn gives the desired result.

A.3. Proof of Lemma 4

Without loss of generality we assume $k_1 \leq k_2$. Following the same argument as in the proof of Theorem 1, we have

$$\begin{aligned}\text{KL}(\mathbb{P}_1, \mathbb{P}_2) &= \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^n \Big(-\frac{1}{2\sigma^2}\left[(y_i - \langle\!\langle \Theta_1,\ X_i \rangle\!\rangle)^2 - (y_i - \langle\!\langle \Theta_2,\ X_i \rangle\!\rangle)^2\right] \Big) \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^n \big[\langle\!\langle \Theta_2,\ X_i \rangle\!\rangle^2 - \langle\!\langle \Theta_1,\ X_i \rangle\!\rangle^2 \\ &\qquad + 2y_i \langle\!\langle \Theta_1,\ X_i \rangle\!\rangle - 2y_i \langle\!\langle \Theta_2,\ X_i \rangle\!\rangle\big] \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^n \Big(\underbrace{\langle\!\langle \Theta_2,\ X_i \rangle\!\rangle - \langle\!\langle \Theta_1,\ X_i \rangle\!\rangle}_{t_i} \Big)^2 = \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^n t_i^2.\end{aligned}$$

We further find that,

$$t_i = \theta_{\min}\left(\sum_{j\in B_1\setminus B_2} X_{ij} - \sum_{j\in B_2\setminus B_1} X_{ij}\right)$$
$$\leq \theta_{\min}\left(\sum_{j\in B_1\Delta B_2} |X_{ij}|\right).$$

Cauchy-Schwarz gives us

$$t_i^2 \leq 2\theta_{\min}^2 k_1 \sum_{j\in B_1\Delta B_2} X_{ij}^2 \leq 2\theta_{\min}^2 k_1,$$

since $\|X_i\|_{\text{fro}}^2 \leq 1$. Combining the results, we have that

$$\text{KL}(\mathbb{P}_1, \mathbb{P}_2) \leq \frac{nk_1\theta_{\min}^2}{\sigma^2}.$$

Together with a similar construction for the case when $k_2 \leq k_1$ we obtain

$$\text{KL}(\mathbb{P}_1, \mathbb{P}_2) \leq \frac{n\min(k_1, k_2)\theta_{\min}^2}{\sigma^2}.$$

A.4. Proof of Lemma 5

The true support B^* is always fully contained in one of the blocks defined in Section 3.2 (see Figure 2). Without loss of generality we assume that the support B^* is fully contained in one block from the first collection. Once we have fixed the collection of blocks Algorithm 1 is invariant to reordering of the blocks, so without loss of generality we can consider the case when B^* is contained in B_{11}. Algorithm 1 on the first collection of blocks proceeds for

$$s_0 := \log\left(\frac{d_1 d_2}{4k_1 k_2}\right)$$

rounds, and in each round we use n_s measurements. It is straightforward to verify that Algorithm 1 errs in the s^{th} round if we have that $w^s < 0$, where

$$\mathbb{P}(w^s < 0) \leq \mathbb{P}\left(N\left(\frac{n_s 2^{(s-1)/2} k_1 k_2 \theta_{\min}}{\sqrt{d_1 d_2}}, n_s\sigma^2\right) < 0\right).$$

A union bound over the rounds gives us a bound on the probability of error as

$$p_e \leq \sum_{s=1}^{s_0} \mathbb{P}[w^s < 0].$$

Recalling the allocation scheme: for $n \geq 2s_0$, $n_s \equiv \lfloor (n - s_0)s2^{-s-1} \rfloor + 1$ and observing that $\sum_{s=1}^{s_0} n_s \leq n$, we have

$$p_e \leq \frac{1}{2} \sum_{s=1}^{s_0} \exp\left(-\frac{n_s 2^s k_1^2 k_2^2 \theta_{\min}^2}{4 d_1 d_2 \sigma^2} \right)$$

using the Gaussian tail bound in Equation (4.3). Since $n_s \geq (n - s_0)s2^{-s-1}$ and $n \geq 2s_0$, we have that $n_s \geq ns2^{-s-2}$. Using this, it is straightforward to verify that if

$$\frac{\theta_{\min}}{\sigma} \geq \sqrt{\frac{16 d_1 d_2}{n k_1^2 k_2^2} \log\left(\frac{1}{2\alpha} + 1 \right)},$$

we have $\mathbb{P}_e \leq \alpha$. We apply this procedure 4 times (once on each collection), and this gives us the desired lemma.

A.5. Proof of Lemma 6

We collect all the rows and columns returned by the 4 runs of Algorithm 1, and condition on the success of Algorithm 1, i.e., we have a set of indices of size at most $(8k_1 \times 8k_2)$, which contains the true support B^*. Without loss of generality we assume these indices are $[8k_1]$ and $[8k_2]$.

Algorithm 2 first identifies one column in B^*. Notice that exactly one of the following columns: $\{1, k_2 + 1, 2k_2 + 1, \ldots, 7k_2 + 1\}$ is contained in B^*. We devote $8\widetilde{n}$ measurements to identify that particular column, where we select $\widetilde{n}$ at the end of lemma to ensure that the total number of measurements used by Algorithm 2 is $n/2$. The procedure is straightforward: measure each column $\widetilde{n}$ times, and pick the one that has the largest total signal.

Verify that for the column in B^* we the total signal we measure is a draw from $N(\sqrt{\frac{k_1}{8}}\theta_{\min}\widetilde{n}, \widetilde{n}\sigma^2)$, while for the other columns we have draws from a Gaussian $N(0, \widetilde{n}\sigma^2)$. Using the Gaussian tail bound from Equation (4.3) it follows that if

$$\frac{\theta_{\min}}{\sigma} \geq \sqrt{\frac{64}{k_1 \widetilde{n}} \log(16/\alpha)},$$

the procedure will identify the column in B^* with probability at least $1 - \alpha/4$.

So far, we have identified a column in B^*. This information tells us which $2k_1$ columns contain B^*. We will use $\widetilde{n}$ more measurements to exactly identify the remaining columns. Rather, than test each of the $2k_2$ columns we will do a binary search. This will require us to test at most $t \equiv 2\lceil \log k_2 \rceil \leq 3 \log k_2$ columns, and we will devote $\widetilde{n}/(3 \log k_2)$ measurements to each column. We will need to threshold these measurements at the threshold:

$$\sqrt{\log\left(\frac{12 \log k_2}{\alpha} \right) \frac{2\widetilde{n}\sigma^2}{3 \log k_2}}$$

and declare that a column belongs to B^* if its average is larger than this threshold. This binary search procedure successfully finds all columns of the block B^* with probability at least $1 - \alpha/4$ if

$$\frac{\theta_{\min}}{\sigma} \geq \sqrt{\frac{32 \log k_2}{\widetilde{n} k_1} \log\left(\frac{3 \log k_2}{\alpha}\right)}$$

We repeat the above procedure to exactly identify the rows of the block B^*. In order to complete the proof of the lemma we need to choose $\widetilde{n}$. Concretely, the choice $\widetilde{n} = n/36$, ensures that the total number of measurements used by Algorithm 2 is $n/2$, and this yields the lemma.

Acknowledgements

We thank Larry Wasserman and Martin Wainwright for several helpful discussions. This research is supported in part by AFOSR under grant FA9550-10-1-0382, NSF under grants IIS-1116458, CCF-1563918, DMS-1149677 and DMS-1713003.

References

[1] S. Aeron, V. Saligrama, and M. Zhao. Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory*, 56(10):5111–5130, 2010. MR2808668

[2] E. Arias-Castro. Detecting a vector based on linear measurements. *Electron. J. Stat.*, 6:547–558, 2012. MR2988419

[3] E. Arias-Castro, E. J. Candés, and M. A. Davenport. On the fundamental limits of adaptive sensing. *IEEE Trans. Inf. Theory*, 59(1):472–481, 2013. MR3008159

[4] E. Arias-Castro, E. J. Candés, and A. Durand. Detection of an anomalous cluster in a network. *Ann. Stat.*, 39(1):278–304, 2011. MR2797847

[5] E. Arias-Castro, E. J. Candés, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Ann. Stat.*, 39(5):2533–2556, 2011. MR2906877

[6] S. Balakrishnan, M. Kolar, A. Rinaldo, and A. Singh. Recovering block-structured activations using compressive measurements. 2012. arXiv:1209.3431. MR2393645

[7] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56(4):1982–2001, 2010. MR2654489

[8] S. Bhamidi, P. S. Dey, and A. B. Nobel. Energy landscape for large average submatrix detection problems in gaussian random matrices. 2012. arXiv:1211.2284. MR3663635

[9] C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

[10] C. Butucea, Y. I. Ingster, and I. Suslina. Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. 2013. arXiv:1303.5647. MR3160567
[11] E. J. Candés and M. A. Davenport. How well can we estimate a sparse vector? *Appl. Comput. Harmon. Anal.*, 34(2):317–323, 2013. MR3008569
[12] E. J. Candés and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.*, 35(6):2313–2351, 2007. MR2382644
[13] E. J. Candés and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Process. Mag.*, 25(2):21–30, 2008.
[14] Y. Caron, P. Makris, and N. Vincent. A method for detecting artificial objects in natural environments. In *16th Int. Conf. Pattern Recogn.*, volume 1, pages 600–603. IEEE, 2002.
[15] R. M. Castro. Adaptive sensing performance lower bounds for sparse signal detection and support estimation. 2012. arXiv:1206.0648. MR3263103
[16] R. M. Castro and E. Tánczos. Adaptive compressed sensing for estimation of structured sparse sets. 2014. arXiv:1410.4593.
[17] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(4):1148–1185, 2012. MR3052407
[18] S. Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015. MR3285604
[19] M. A. Davenport and E. Arias-Castro. Compressive binary search. 2012. arXiv:1202.0937.
[20] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006. MR2241189
[21] M. F. Duarte, M. A. Davenport, M. J. Wainwright, and R. G. Baraniuk. Sparse signal detection from incoherent projections. In *Proc. IEEE Int. Conf. Acoustics Speed and Signal Processing*, pages III–305–III–308. 2006. MR2451158
[22] C. F. F. C. Filho, R. de Oliveira Melo, and M. G. F. Costa. *Detecting Natural Gas Leaks Using Digital Images and Novelty Filters*, pages 242–249. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
[23] G. M. Foody and A. Mathur. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(6):1335–1343, 2004.
[24] M. Golbabaee and P. Vandergheynst. Compressed sensing of simultaneous low-rank and joint-sparse matrices. 2012. arXiv:1211.5058.
[25] J. D. Haupt, R. G. Baraniuk, R. M. Castro, and R. D. Nowak. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Proc. 43rd Asilomar Conf. Signals, Systems and Computers*, pages 1551–1555. IEEE, 2009.
[26] J. D. Haupt and R. D. Nowak. Compressive sampling for signal detection. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, volume 3, pages III–1509–III–1512. IEEE, 2007.
[27] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *J. Mach. Learn. Res.*, 12:3371–3412, 2011. MR2877603

[28] Y. I. Ingster, A. B. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Electron. J. Stat.*, 4:1476–1526, 2010. MR2747131

[29] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 909–917. 2011.

[30] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.*, 39(5):2302–2329, 2011. MR2906869

[31] A. Krishnamurthy, J. Sharpnack, and A. Singh. Recovering graph-structured activations using adaptive compressive measurements. 2013. arXiv:1305.0213.

[32] Y. Ma, D. J. Sutherland, R. Garnett, and J. G. Schneider. Active pointillistic pattern search. In *AISTATS*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

[33] M. L. Malloy and R. D. Nowak. Near-optimal adaptive compressed sensing. In *Proc. 46th Asilomar Conf. Signals, Systems and Computers*, pages 1935–1939. IEEE, 2012. MR3225946

[34] M. L. Malloy and R. D. Nowak. Near-optimal compressive binary search. 2012. arXiv:1203.1804.

[35] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Stat.*, 39(2):1069–1097, 2011. MR2816348

[36] G. Reeves and M. Gastpar. The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Trans. Inf. Theory*, 58(5):3065–3092, 2012. MR2952533

[37] E. Richard, P.-A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. In *Proc. 29th Int. Conf. Mach. Learn.*, pages 1351–1358. 2012.

[38] A. Soni and J. D. Haupt. Efficient adaptive compressive sensing using sparse hierarchical learned dictionaries. In *Proc. 45th Conf. Signals, Systems and Computers*, pages 1250–1254. IEEE, 2011.

[39] A. Soni and J. D. Haupt. On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. 2013. arXiv:1306.4391. MR3150916

[40] X. Sun and A. B. Nobel. On the maximal size of large-average and ANOVA-fit submatrices in a gaussian random matrix. *Bernoulli*, 19(1):275–294, 2013. MR3019495

[41] E. Tánczos and R. M. Castro. Adaptive sensing for estimation of structured sparse signals. 2013. arXiv:1311.7118. MR3332997

[42] A. B. Tsybakov. *Introduction To Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. MR2724359

[43] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.

[44] M. M. Wagner, F. C. Tsui, J. U. Espino, V. M. Dato, D. F. Sittig, R. A. Caruana, L. F. McGinnis, D. W. Deerfield, M. J. Druzdzel, and D. B. Fridsma. The emerging science of very early detection of disease outbreaks. *J Public Health Manag Pract*, 7(6):51–59, 2001.

[45] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory*, 55(12):5728–5741, 2009. MR2597190

[46] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202, 2009. MR2729873

[47] S. Yoon, C. Nardini, L. Benini, and G. De Micheli. Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2(4):339–354, 2005.