arXiv: 0000.0000

RECOVERING BLOCK-STRUCTURED ACTIVATIONS USING COMPRESSIVE MEASUREMENTS*

By Sivaraman Balakrishnan, Mladen Kolar, Alessandro Rinaldo and Aarti Singh

Carnegie Mellon University

We consider the problem of detection and localization of a small block of weak activation in a large matrix, from a small number of noisy, possibly adaptive, compressive (linear) measurements. This is closely related to the problem of compressed sensing, where the task is to estimate a sparse vector using a small number of linear measurements. However, contrary to results in compressed sensing, where it has been shown that neither adaptivity nor contiguous structure help much, we show that in our problem the magnitude of the weakest signals one can reliably localize is strongly influenced by both structure and the ability to choose measurements adaptively. We derive tight upper and lower bounds for the detection and estimation problems, under both adaptive and non-adaptive measurement schemes. We characterize the precise tradeoffs between the various problem parameters, the signal strength and the number of measurements required to reliably detect and localize the block of activation.

1. Introduction. Compressive measurements provide a very efficient means of recovering data vectors that are sparse in some basis or frame. Specifically, several papers, including Candès and Tao (2006), Donoho (2006), Candès and Tao (2007), Candès and Wakin (2008), and Wainwright (2009a) have shown that it is possible to recover a k-sparse vector in n dimensions using only $k \log n$ compressive measurements, instead of measuring all of the n coordinates. Motivated by this line of research, there have been recent attempts (Baraniuk et al., 2010, Soni and Haupt, 2011) at characterizing the number of compressive measurements needed to recover vectors that are endowed with some *structure* in addition to sparsity. Yet another extension of the compressed sensing framework has been to attempt to recover vectors from few possibly *adaptive* compressed measurements, where sub-

*We would like to thank Larry Wasserman for his ideas, indispensable advice and wise guidance. This research is supported in part by AFOSR under grant FA9550-10-1-0382, NSF under grant IIS-1116458 and NSF CAREER grant DMS 1149677.

AMS 2000 subject classifications: Primary 62F03, 62F12; secondary 62F12, 62L12

Keywords and phrases: adaptive procedures, compressive measurements, large average submatrix, signal detection, signal localization, structured Normal means

sequent measurements are designed based on past observations (see, e.g., Candès and Davenport, 2011). Finally, there has also been work on detection, instead of recovery, of sparse vectors from compressive measurements (Arias-Castro, 2012). However, almost all of this work has been focused on recovery or detection of (structured or unstructured) sparse data *vectors* from (passive or adaptive) compressed measurements.

In this paper, we extend the compressed sensing paradigm to handle data *matrices*. In the unstructured case, the treatment of data matrices is exactly equivalent to the treatment of data vectors. The setting where data matrices are distinct from data vectors is when the sparsity pattern is structured in a way that reflects some coupling between the rows and columns. We consider one such setup where there is a sub-matrix or block of activation embedded in the data matrix. This is a natural model for several real-world activations such as when we have a group of genes (belonging to a common pathway for instance) co-expressed under the influence of a set of similar drugs (Yoon et al., 2005), when we have groups of patients exhibiting similar symptoms (Moore et al., 2010), when we have sets of malware with similar signatures (Jang et al., 2011), etc. However, in many of these applications, it is difficult to measure, compute or store all the entries of the data matrix. For example, measuring expression levels of all genes under all possible drugs is expensive, or recording the signatures of each individual malware is computationally demanding as it might require stepping through the entire malware code. However, if we have access to linear combinations of matrix entries (i.e. compressive measurements) such as combined expression of multiple genes under the influence of multiple drugs then we might need to only make and store few such measurements, while still being able to infer the existence or location of the activated block of the data matrix. Thus, the goal is to detect or recover the activated block (set of co-expressed genes and drugs or malwares with similar signatures) using only few compressive measurements of the data matrix, instead of observing the entire data matrix directly. We consider both the passive (non-adaptive) and active (adaptive) measurements. The non-adaptive measurements are random or pre-specified linear combinations of matrix entries. In other cases, such as mixing drugs, we might be able to adapt the measurement process and sequentially design linear combinations that are more informative.

Summary of our contributions. Using information theoretic tools, we establish *lower bounds* on the minimum number of compressive measurements and the weakest signal-to-noise ratio (SNR) needed to detect the presence of an activated block of positive activation, as well as to localize the activated block, using both non-adaptive and adaptive measurements.

TABLE 1
Summary of main findings under the assumption that $n_1 = n_2 = n$ and $k_1 = k_2 = k$, where the size of the matrix is $n \times n$ and the size of the activation block is $k \times k$. The number of measurements is m and μ/σ represents SNR per element of the activated block.

	Detection	Localization	
Passive	$\frac{\mu}{\sigma} \asymp \frac{1}{\sqrt{m}} \frac{n}{k^2}$	$\frac{\mu}{\sigma} \asymp \frac{1}{\sqrt{m}} \frac{n}{\sqrt{k}}$	Theorems 3 and 4
Active	Theorems 1 and 2	$\frac{\mu}{\sigma} \asymp \frac{1}{\sqrt{m}} \max\left(\frac{n}{k^2}, \frac{1}{\sqrt{k}}\right)$	Theorems 5 and 6

We also demonstrate minimax optimal *upper bounds* through detectors and estimators that can guarantee consistent detection and recovery of weak block-structured activations using few non-adaptive and adaptive compressive measurements.

Our results indicate that adaptivity and structure play a key role and provide significant improvements over non-adaptive and unstructured cases for recovery of the activated block in the data matrix setting. This is unlike the vector case where contiguous structure and adaptivity have been shown to provide minor, if any, improvement (Candès and Davenport, 2011).

In our setting we take compressive measurements of a data *matrix* of size $(n_1 \times n_2)$, the activated block is of size $(k_1 \times k_2)$, with minimum SNR per entry of μ/σ, and we have a budget of m compressive measurements with each measurement matrix constrained to have unit Frobenius norm.

Table 1 describes our main findings (assuming $n_1 = n_2 = n$ and $k_1 = k_2 = k$ and paraphrasing for clarity) and compare the scalings under which passive and active, detection and localization are possible.

For detection, akin to the vector setting, structure and adaptivity play no role. The structured data matrix setting requires an SNR scaling as $\sqrt{n_1 n_2/(m k_1^2 k_2^2)}$ for both non-adaptive and adaptive cases, which is same as the SNR needed to detect a $k_1 k_2$ sparse non-negative vector of length $n_1 n_2$ as demonstrated in Arias-Castro (2012). Thus, the structure of the activation pattern as well as the power of adaptivity offer no advantage in the detection problem.

For localization of the activated block, the structured data matrix setting requires an SNR scaling as $\sqrt{n_1 n_2/(m \min(k_1, k_2))}$ using non-adaptive compressive measurements. In contrast, the unstructured setting requires a higher SNR of $\sqrt{n_1 n_2 \log(n_1 n_2)/m}$ where $m \geq k_1 k_2 \log(n_1 n_2)$ as demonstrated in Wainwright (2009b). Structure, without adaptivity already yields a factor of $\sqrt{k}$ reduction in the smallest SNR that still allows for reliable localization. Moreover, adaptivity in the compressive measurement design yields further improvements. With adaptive measurements, identifying the activated block requires a much weaker SNR of $\max(\sqrt{n_1 n_2/(m k_1^2 k_2^2)}, \sqrt{1/(m \min(k_1, k_2))})$ for the weakest entry in the data matrix. For the sparse vector case, Arias-Castro et al. (2011) showed that adaptive compressive measurements cannot recover the non-zero locations if the SNR is smaller than $\sqrt{n_1 n_2/m}$. A matching upper bound was provided using compressive binary search in Davenport and Arias-Castro (2012) and Malloy and Nowak (2012) for recovering the location of a single non-zero entry in the vector. Thus, exploiting structure of the activations and designing adaptive linear measurements can both yield significant gains if the activation corresponds to a block in a data matrix.

Related Work. Our work builds on a number of fairly recent contributions on detection and recovery of a sparse and weak unstructured signal by adaptive compressive measurements. In Arias-Castro et al. (2011), the authors show that, in the linear regression setting, the adaptive compressive scheme offers improvements over the passive scheme which, in terms of MSE, are limited to a $\log(n)$ factor. The authors also provide a general proof strategy for minimax analysis under adaptive measurements. Arias-Castro (2012) further applies this strategy to the problem of detection of an unstructured sparse and weak vector signal under compressive adaptive measurements. Malloy and Nowak (2012) shows that a compressive version of standard binary search achieves minimax performance for localization in a one-sparse vector. The work of Wainwright (2009b) which is based on analyzing the performance of an exhaustive search procedure under passive measurements, is relevant to our analysis of passive localization. Our analysis provides a generalization of these results to the case of a *structured* and weak signal embedded as a small contiguous block in a large matrix.

While we focus on detection and localization of the activation in this paper, some other papers have considered estimation of sparse vectors in the mean square error (MSE) sense using adaptive compressive measurements. For example, Candès and Davenport (2011) establishes fundamental lower bounds on the MSE in a linear regression framework, while Haupt et al. (2009) demonstrates upper bounds using compressive distilled sensing. Some other papers (Baraniuk et al., 2010, Soni and Haupt, 2011) have considered different forms of structured sparsity in the vector setting, e.g. if the non-zero locations in a data vector form non-overlapping or partially-overlapping groups or are tree-structured. Finally, Negahban and Wainwright (2011) and Koltchinskii et al. (2011) have considered a measurement model identical to ours in the setting of low-rank matrix completion, but in that setting the matrix under consideration is not assumed to be a structured sparse matrix and the theoretical guarantees are with respect to the Frobenius norm.

The rest of this paper is organized as follows. We describe the problem set up and notation in Section 2. We study the detection problem in Section 3, for both adaptive and non-adaptive schemes. Section 4 is devoted to the non-adaptive localization, while Section 5 is focused on adaptive localization. Finally, in Section 6 we present and discuss some simulations that support our findings. The proofs are given in the Appendix.

2. Preliminaries. Let $A \in \mathbb{R}^{n_1 \times n_2}$ be a signal matrix with unknown entries that we would like to recover. We are interested in a highly *structured* setting where a *contiguous* block of the matrix A of size $(k_1 \times k_2)$ has entries all equal to $\mu > 0$, while all the other elements of A are equal to zero. Define the set of contiguous blocks,

(2.1)
$$\mathcal{B} = \{I_r \times I_c : I_r \text{ and } I_c \text{ are contiguous subsets of } [n_1] \text{ and } [n_2]^1, |I_r| = k_1, |I_c| = k_2\}.$$

Then $A = (a_{ij})$ with $a_{ij} = \mu \, \mathbb{I}\{(i,j) \in B^*\}$ for some (unknown) $B^* \in \mathcal{B}$. All of our results extend to the case when the activation is not constant on B^*, with $\min_{(i,j)\in B^*} a_{ij}$ replacing μ in all our results.

We consider the following observation model under which m noisy linear measurements of A are available

$$y_i = \mathrm{tr}(\mathrm{AX_i}) + \epsilon_i, \quad i = 1, \ldots, m, \tag{2.2}$$

where $\epsilon_1, \ldots, \epsilon_m \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $\sigma > 0$ known, and the sensing matrices $(X_i)_i$ satisfy either $\|X_i\|_F \leq 1$ or $\mathbb{E}\|X_i\|_F^2 = 1$.

Under the observation model in Eq. (2.2), we study two tasks: (1) detecting whether a contiguous block of positive signal exists in A and (2) identifying the block B^*, that is, the localization of B^*. We develop efficient algorithms for these two tasks that provably require the smallest number of measurements, as explained below. The algorithms are designed for one of two measurement schemes: (1) the measurement scheme can be implemented in an adaptive or sequential fashion, that is, actively, by letting each X_i to be a (possibly randomized) function of $(y_j, X_j)_{j\in[i-1]}$, and (2) the measurement matrices are chosen all at once, that is, passively.

Detection. The detection problem concerns checking whether a positive contiguous block exists in A. As we will show later, we can detect presence of a contiguous block with much smaller number of measurements than is required for localizing its position. Therefore, solving the detection problem before trying to localize the block is often important. Formally, detection is

[1] We use $[n]$ to denote the set $\{1, \ldots, n\}$

a hypothesis testing problem with a composite alternative of the form

$$\begin{array}{ll} H_0: & A = 0_{n_1 \times n_2} \\ H_1: & A = (a_{ij}) \text{ with } a_{ij} = \mu\, \mathbb{I}_{\{(i,j)\in B\}},\ B \in \mathcal{B}. \end{array} \tag{2.3}$$

A test T is a measurable function of the observations and the measurements matrices $(y_i, X_i)_{i\in[m]}$, which takes values in $\{0,1\}$, $T = 1$ if the null hypothesis is rejected and $T = 0$ otherwise. For any test T, we define its risk as

$$R(T) \equiv \mathbb{P}_0\left[T\big((y_i, X_i)_{i\in[m]}\big) = 1\right] + \max_{B\in\mathcal{B}} \mathbb{P}_B\left[T\big((y_i, X_i)_{i\in[m]}\big) = 0\right],$$

where $\mathbb{P}_0$ and $\mathbb{P}_B$ denote the joint probability distributions of $\big((y_i, X_i)_{i\in[m]}\big)$ under the null hypothesis and when the activation pattern is B, respectively. The risk $R(T)$ measures the maximal sum of type I and type II errors over the set of alternatives. The overall difficulty of the detection problem is quantified by the *minimax risk* $R \equiv \inf_T R(T)$, where the infimum is taken over all tests. For a sufficiently small SNR, the minimax risk is bounded away from zero by a large constant, which implies that no test can distinguish H_0 from H_1. We precisely characterize the boundary for SNR below which no test can distinguish H_0 and H_1.

Localization. The localization problem concerns recovery of the true activation pattern B^*. Let Ψ be an estimator of B^*, with the risk, corresponding to a 0/1 loss, given by

$$R(\Psi) = \max_{B\in\mathcal{B}} P_B\left[\Psi\big((y_i, X_i)_{i\in[m]}\big) \neq B\right],$$

while the *minimax risk* of the localization problem is the minimal risk over all such estimators Ψ. Like in the detection task, the minimax risk specifies the minimal risk of any localization procedure. By standard arguments, the evaluation of the minimax localization risk also proceeds by first reducing the localization problem to a hypothesis testing problem (see, e.g., Tsybakov, 2009, for details).

Below we will provide a sharp characterization, through information theoretic lower bounds and tractable estimators, of the minimax detection and localizations risks as functions of tuples of $(n_1, n_2, k_1, k_2, m, \mu, \sigma)$ and for both the active and passive sampling schemes. Our results identify precisely both the minimal SNR given a budget of m possibly adaptive measurements, and the minimal number of measurements m for a given SNR in order to achieve successful detection and localization.

Along with a careful and detailed minimax analysis, we also describe procedures for detection and localization in both the active and passive case whose risks match the minimax rates.

3. Detection of contiguous blocks.

In this section, we provide a sharp characterization of the minimax detection risk.

3.1. *Lower bound.*

The following theorem gives a lower bound on the SNR needed to distinguish H_0 and H_1.

THEOREM 1. *Fix any $0 < \alpha < 1$. Based on m (possibly adaptive) measurements, if $\mu < \mu_{\min}$, where*

$$\mu_{\min} := \sigma(1-\alpha)\sqrt{\frac{16(n_1-k_1)(n_2-k_2)}{mk_1^2k_2^2}},$$

then any test to distinguish H_0 from H_1, defined in Eq. (2.3), has risk at least α.

The result of Theorem 1 can be interpreted as follows: whatever the test T and the risk level α are, there exists $A = (a_{ij})$ with $a_{ij} = \mu\,\mathbb{I}\{(i,j) \in B^*\}$, $\mu < \mu_{\min}$, such that $R(T) \geq \alpha$. This gives a lower bound on the minimax risk as $\inf_T R(T) \geq \alpha$.

The lower bound on possibly *adaptive* procedures is established by analyzing the risk of the (optimal) likelihood ratio test under a certain prior on the alternatives. Careful modifications of standard arguments are necessary to account for adaptivity. We closely follow the approach of Arias-Castro Arias-Castro (2012) who established the analogue of Theorem 1 in the vector setting.

3.2. *Upper bound.*

We now discuss the sharpness of the result established in the previous section. We choose the sensing matrices passively as $X_i = (n_1n_2)^{-1/2}\mathbf{1}_{n_1}\mathbf{1}'_{n_2}$ and consider the following test

$$T\big((y_i)_{i\in[m]}\big) = \mathbb{I}\Big\{\sum_i y_i > \sigma\sqrt{2m\log(\alpha^{-1})}\Big\}. \tag{3.1}$$

THEOREM 2. *If $\mu > \sigma\sqrt{\frac{8n_1n_2\log(\alpha^{-1})}{mk_1^2k_2^2}}$ then $R(T) \leq \alpha$, where T is the test defined in Eq. (3.1).*

Results of Theorem 1 and Theorem 2 establish that the minimax rate for detection under the model in Eq. (2.2) is $\mu \asymp \sigma(k_1k_2)^{-1}\sqrt{m^{-1}n_1n_2}$, under the (mild) assumption that $k_1 \leq cn_1$ and $k_2 \leq cn_2$ for any constant $0 < c < 1$. It is worth pointing out that the structure of the activation pattern *does not* play any role in the minimax detection problem. We will contrast this to the localization problem below. Furthermore, the procedure that achieves the adaptive lower bound (upto constants) is non-adaptive, indicating that adaptivity can not help much in the detection problem.

4. Localization from passive measurements.

In this section, we address the problem of estimating a contiguous block of activation B^* from noisy linear measurement in Eq. (2.2), when the measurement matrices $(X_i)_{i\in[m]}$ are independent with i.i.d. entries $x_{i,ab} \sim \mathcal{N}(0,(n_1 n_2)^{-1})$. The variance of the elements is set so that $\mathbb{E}||X_i||_F^2 = 1$.

4.1. *Lower bound.*

The following theorem gives a lower bound on the SNR needed for any procedure to localize B^*.

THEOREM 3. *There exist two positive constant $C, C' > 0$ independent of the problem parameters (k_1, k_2, n_1, n_2), such that if $\mu < \mu_{\min}^{\text{loc}}$, where*

$$\mu_{\min}^{\text{loc}} := C\sigma\sqrt{\frac{n_1 n_2}{m}\max\left(\frac{1}{\min(k_1,k_2)}, \frac{\log\max(n_1-k_1, n_2-k_2)}{k_1 k_2}\right)},$$

then $\inf_\Psi R(\Psi) \geq C' > 0$ *as* $n \to \infty$.

The proof is based on a standard technique described in Chapter 2.6 of Tsybakov (2009). We start by identifying a subset of matrices from $\mathcal{A}$ which are hard to distinguish. Once a suitable finite set is identified, tools for establishing lower bounds on the error in multiple-hypothesis testing can be directly applied. These tools only require computing the Kullback-Leibler (KL) divergence between the induced distributions, which in our case are two multivariate normal distributions.

The two terms in the lower bound feature two aspects of our construction, the first term arises from considering two matrices that overlap considerably, while the second term arises from considering matrices that do not overlap at all of which there are possibly a very large number. These constructions and calculations are described in detail in the Appendix.

4.2. *Upper bound.*

We will investigate a procedure that searches over all contiguous blocks of size $(k_1 \times k_2)$ defined in Eq. (2.1) and outputs one that minimizes the squared error. Define the loss function $f : \mathcal{B} \mapsto \mathbb{R}$ as

$$f(B) := \min_\mu \sum_{i\in n}\Big(\mu \sum_{(a,b)\in\mathcal{B}} x_{i,ab} - y_i\Big)^2. \qquad (4.1)$$

Then the estimated block $\widehat{B}$ is defined as

$$\widehat{B} := \operatorname*{argmin}_{B \in \mathcal{B}} f(B). \tag{4.2}$$

Note that the minimization problem above requires solving $O(n_1 n_2)$ univariate regression problems and can be implemented efficiently for reasonably large matrices.

The following results characterizes the SNR needed for $\widehat{B}$ to correctly identify B^*.

THEOREM 4. *There exists a positive constant $C > 0$ independent of the problem parameters (k_1, k_2, n_1, n_2), such that if*

$$\mu \geq C\sigma \sqrt{\frac{n_1 n_2}{m} \log(2/\alpha) \max\left(\frac{\log \max(k_1, k_2)}{\min(k_1, k_2)}, \frac{\log \max(n_1 - k_1, n_2 - k_2)}{k_1 k_2}\right)},$$

then $R(\widehat{B}) \leq \alpha$, where $\widehat{B}$ is defined in Eq. (4.2).

Comparing to the lower bound in Theorem 3, we observe that the procedure outlined in this section achieves the lower bound up to constants and a $\log k$ factor. Under the scaling $\min(k_1, k_2) \geq \log \max(n_1 - k_1, n_2 - k_2)$, we obtain that the *passive* minimax rate for localization of the active blocks B^* is $\mu \asymp \widetilde{O}\big(\sigma\sqrt{(m \min(k_1, k_2))^{-1} n_1 n_2}\big)$. This establishes that the SNR needed for passive localization is considerably larger than the bound we saw earlier for passive detection. This should be contrasted to the normal means problem, where the bounds for localization and detection differ only in constants (Donoho and Jin, 2004).

The block structure of the activation allows us, even in the passive setting, to localize much weaker signals. A straightforward adaptation of results on the LASSO (Wainwright, 2009a) suggest that if the non-zero entries are spread out (say at random) then we would require $\mu \asymp \widetilde{O}\left(\sigma\sqrt{\frac{n_1 n_2}{m}}\right)$ for localization.

5. Localization from active measurements.

In this section, we study localization of B^* using adaptive procedures, that is, the measurement matrix X_i may be a function of $(y_j, X_j)_{j \in [i-1]}$.

5.1. *Lower bound.*

A lower bound on the SNR needed for any active procedure to localize B^* is given.

THEOREM 5. *Fix any $0 < \alpha < 1$. Given m adaptively chosen measurements, if $\mu < \mu_{\min}^{\text{loc,active}}$, where*

$$\mu_{\min}^{\text{loc,active}} := \sigma(1-\alpha) \max\left(\sqrt{\frac{2\max((n_1-k_1)(n_2/2-k_2),(n_1/2-k_1)(n_2-k_2))}{mk_1^2k_2^2}}, \sqrt{\frac{8}{m\min(k_1,k_2)}} \right),$$

then $\inf_\Psi R(\Psi) \geq \alpha$.

The proof is based on an information theoretic arguments applied to specific pairs of hypotheses that are hard to distinguish. The two terms in the lower bound reflect the two sources of hardness of the problem of exactly localizing the block of activation. The first term reflects the hardness of approximately localizing the block of activation. This term grows at the same rate as the detection lower bound, and its proof is similar. Given a coarse localization of the block we still need to exactly localize the block. The hardness of this problem gives rise to the second term in the lower bound. The term is independent of n_1 and n_2 but has a considerably worse dependence on k_1 and k_2.

5.2. *Upper bound.* The upper bound is established by analyzing the procedures described in Algorithms 1 and 2 for approximate and exact localization. Algorithm 1 is used to approximately located the activation block, that is, it locates a $2s \times 2s$ blocks that contains the activation block with high probability. The algorithm essentially performs the compressive binary search on a collection of non-overlapping blocks that partition the signal matrix. It is run on two collections, $\mathcal{D}_1$ and $\mathcal{D}_2$, defined as

$$\mathcal{D}_1 \equiv \{B_{11} = [1,\ldots,2k_1][1,\ldots,2k_2] \cup B_{12} = [2k_1+1,\ldots,4k_1][1,\ldots,2k_2] \cup \ldots$$
$$\ldots \cup B_{1n_1n_2/4k_1k_2} = [n_1-2k_1,\ldots,n_1][n_2-2k_2,\ldots,n_2]\}$$

and

$$\mathcal{D}_2 \equiv \{B_{21} = [k_1,\ldots,3k_1][k_2,\ldots,3k_2] \cup B_{22} = [3k_1+1,\ldots,5k_1][k_2,\ldots,3k_2] \cup \ldots$$
$$\cup \ldots B_{2n_1n_2/4k_1k_2} = [n_1-k_1,...,n_1,1,\ldots,k_1][n_2-k_2,...,n_2,1,\ldots,k_2]\}.$$

Notice, that one of these collections must contain a block with the *full* block of activation. Algorithm 1 applied twice returns two blocks, one of which as we show has the desired block with high probability.

Algorithm 2 is used next to precisely locate the activation block within one of the two coarser blocks identified by Algorithm 1. Algorithm 2 is a

Algorithm 1 Approximate localization

input Measurement budget $m \geq \log p$, (dyadic) ordered collection of size p of blocks $\mathcal{D}$ of size $(u_1 \times u_2)$

Initial support: $J_0^{(1)} \equiv \{1, \ldots, p\}$, $s_0 \equiv \log p$

For each s in $1, \ldots, \log_2 p$

1. Allocate: $m_s \equiv \lfloor (m - s_0) s 2^{-s-1} \rfloor + 1$
2. Split: $J_1^{(s)}$ and $J_2^{(s)}$, left and right half collections of blocks of $J_0^{(s)}$
3. Sensing matrix: $X_s = \sqrt{\frac{2^{-(s_0 - s + 1)}}{u_1 u_2}}$ on $J_1^{(s)}$, $X_s = -\sqrt{\frac{2^{-(s_0 - s + 1)}}{u_1 u_2}}$ on $J_2^{(s)}$ and 0 otherwise.
4. Measure: $y_i^{(s)} = \mathrm{tr}(A X_s) + z_i^{(s)}$, for $i \in [1, \ldots, m_s]$
5. Update support: $J_0^{(s+1)} = J_1^{(s)}$ if $\sum_{i=1}^{m_s} y_i^{(s)} > 0$ and $J_0^{(s+1)} = J_2^{(s)}$ otherwise

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

modified compressive binary search procedure that is used to quickly zoom in on the active rows and columns within a larger block.

The following theorem states that Algorithm 1 and Algorithm 2 succeed in localization of the active block with high probability if SNR is large enough.

THEOREM 6. *If*

$$\mu \geq \sigma \sqrt{\log(1/\alpha)} \widetilde{O} \left(\max \left(\sqrt{\frac{n_1 n_2}{m k_1^2 k_2^2}}, \sqrt{\frac{1}{\min(k_1, k_2) m}} \right) \right)$$

and $m \geq 3 \log(n_1 n_2)$ *then* $\inf_\Psi R(\Psi) \leq \alpha$.

The $\widetilde{O}$ hides a $\sqrt{\log \max(k_1, k_2)}$ factor, and our upper bound matches the lower bound up to this factor. It is worth noting that for small activation blocks (when the first term dominates) our active localization procedure achieves the *detection* limits. This is the best result we could hope for. For larger activation blocks, the lower bound indicates that *no* procedure can achieve the detection rate. The active procedure still remains significantly more efficient than the passive one, and even in this case is able to detect signals that are weaker by a (large) $\sqrt{n_1 n_2}$ factor. This is not the case for compressed sensing with vectors as shown in Arias-Castro et al. (2011). The great potential for gains from adaptive measurements is clearly seen in our model which captures the fundamental interplay between *structure* and *adaptivity*.

Algorithm 2 Exact localization

input Measurement budget $5m$, a sub-matrix $B \in \mathbb{R}^{4k_1 \times 4k_2}$

Measure: $y_i^c = (4k_1)^{-1/2} \sum_{l=1}^{4k_1} B_{lc} + z_i^c$, for $i = \{1, \ldots, m\}$ and $c \in \{1, k_2+1, 2k_2+1, 3k_2+1\}$

Let $l = \text{argmax}_c \sum_{i=1}^{m} y_i^c$

Let $r = l + k_2$

Let $m_b = \lfloor \frac{m}{3 \log_2 k_2} \rfloor$

While $r - l > 1$

1. Let $c = \lfloor \frac{r+l}{2} \rfloor$
2. Measure $y_i^c = (4k_1)^{-1/2} \sum_{l=1}^{4k_1} B_{lc} + z_i^c$ for $i = \{1, \ldots, m_b\}$
3. If $\sum_{i=1}^{m_b} y_i^c \geq \tau$, then $l = c$, otherwise $r = c$

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

6. Experiments.

In this section, we perform a set of simulation studies to illustrate finite sample performance of the proposed procedures. We let $n_1 = n_2 = n$ and $k_1 = k_2 = k$. Theorem 4 and Theorem 6 characterize the SNR needed for the passive and active 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 appropriately rescaled SNR and showing that the curves for different values of n and k line up.

Experiment 1. Figure 1 shows the probability of successful localization of B^* using $\widehat{B}$ defined in Eq. (4.2) plotted against $n^{-1}\sqrt{km} * \text{SNR}$, where the number of measurements $m = 100$. Each plot in Figure 1 represents different relationship between k and n; in the first plot, $k = \mathcal{O}(\log n)$, in the second $k = \mathcal{O}(\sqrt{n})$, while in the third plot $k = \mathcal{O}(n)$. The dashed vertical line denotes the threshold position for the scaled SNR at which the probability of success is larger than 0.95. We observe that irrespective of the problem size and the relationship between n and k, Theorem 4 tightly characterizes the minimum SNR needed for successful identification.

Experiment 2. Figure 2 shows the probability of successful localization of B^* using the procedure outlined in Section 5.2., with $m = 500$ adaptively chosen measurements, plotted against the scaled SNR. The SNR is scaled by $n^{-1}\sqrt{m}k^2$ in the first two plots where $k = \mathcal{O}(\log n)$ and $k = \mathcal{O}(\sqrt{n})$ respectively, while in the third plot the SNR is scaled by $\sqrt{mk/\log k}$ as $k = \mathcal{O}(n)$. The dashed vertical line denotes the threshold position for the scaled SNR at which the probability of success is larger than 0.95. We observe that Theorem 6 sharply characterizes the minimum SNR needed for successful

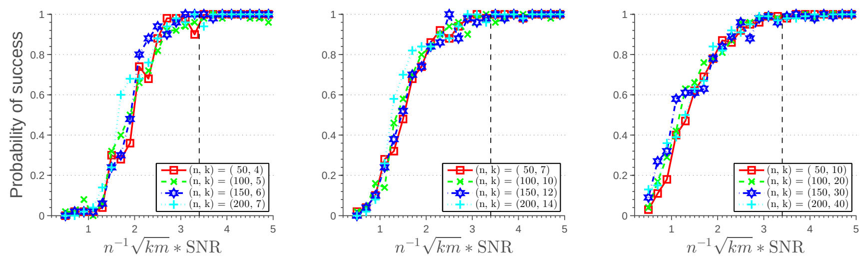

Fig 1. *Probability of success with passive measurements (averaged over 100 simulation runs).*

identification.

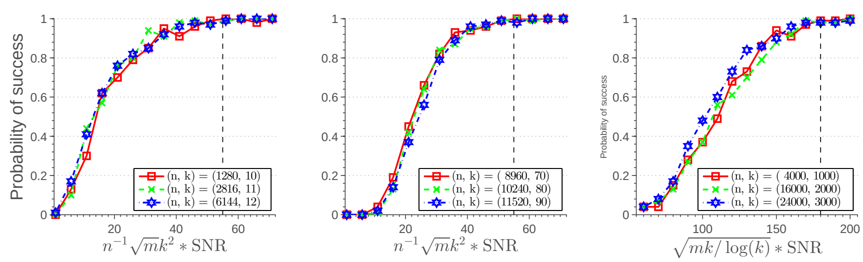

Fig 2. *Probability of success with adaptively chosen measurements (averaged over 100 simulation runs).*

7. Discussion. In this paper, we establish the fundamental limits for the problem of detecting and localizing a block of weak activation in a data matrix from either adaptive or non-adaptive compressive measurements. Our bounds precisely characterize the tradeoff between signal-to-noise ratio, size of matrix, size of sub-matrix and number of measurements. We also demonstrate constructive computationally efficient procedures that achieve these bounds. Contrary to recent results for sparse vectors which demonstrate that contiguous structure for the activation and the ability to choose measurements adaptively play a negligible role in detection and localization, our results indicate that both the block-structure of the activation and adaptive measurement design significantly improve the localization performance for data matrices. An intuitive explanation for why adaptive sampling helps in the structured case is that in this case it is possible to quickly focus the sampling using a compressive binary search procedure, and then exploit the structure for exact localization. In the unstructured case however the signal can be spread out and the adaptive procedure has no way to rule out candidate locations quickly and has to repeatedly measure essentially all locations.

In this paper, we assumed that an ordering of rows and columns of the data matrix is available. Such an ordering may be obtained by a preprocessing step that clusters the rows and columns of the matrix. However, the general problem of recovering an activated block within a randomly permuted data matrix, commonly known as biclustering, is also important. The biclustering problem can be harder than the un-permuted setting as established in Kolar et al. (2011), at least when all the matrix entries can be directly observed and we hope to address its compressive analog in future work.

One important open question, that remains unsolved, is the problem of finding the size of the activation block in a data dependent way. At the moment we are not aware of procedures that can localize the activation block at the minimax SNR without the knowledge of its size. Butucea and Ingster (2011) propose test procedures, under a slightly different model, for detection of the activation block that do not require the knowledge of the size, but work with a collection of sizes that contain the true size. However, the price for being agnostic to the size is reflected in the established rates, which reflect the difficulty of detecting the hardest activation block in the collection. Therefore, even for the problem of detection, adaptation to the size is an open problem.

APPENDIX A: PROOFS OF MAIN RESULTS

In this appendix, we collect proofs of the results stated in the paper. Throughout the proofs, we will denote $c_1, c_2, \dots$ positive constants that may change their value from line to line.

A.1. Proof of Theorem 1. We lower bound the Bayes risk of any test T. Recall, the null and alternate hypothesis, defined in Eq. (2.3),

$$\begin{aligned} H_0: &\quad A = 0_{n_1 \times n_2} \\ H_1: &\quad A = (a_{ij}) \text{ with } a_{ij} = \mu \, \mathbb{I}_{\{(i,j) \in B\}}, \ B \in \mathcal{B}. \end{aligned}$$

We will consider a uniform prior over the alternatives π, and bound the average risk

$$R_\pi(T) = \mathbb{P}_0[T = 1] + \mathbb{E}_{A \sim \pi} \mathbb{P}_A[T = 0],$$

which provides a lower bound on the worst case risk of T.

Under the prior π, the hypothesis testing becomes to distinguish

$$\begin{aligned} H_0: &\quad A = 0_{n_1 \times n_2} \\ H_1: &\quad A = (a_{ij}) \text{ with } a_{ij} = \mathbb{E}_{B \sim \pi} \mu \, \mathbb{I}_{\{(i,j) \in B\}} \, . \end{aligned}$$

Both H_0 and H_1 are simple and the likelihood ratio test is optimal by the Neyman-Pearson lemma. The likelihood ratio is

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

where the second equality follows by decomposing the probabilities by the chain rule and observing that $P_0[X_i|(y_j, X_j)_{j\in[i-1]}] = P_A[X_i|(y_j, X_j)_{j\in[i-1]}]$, since the sampling strategy (whether active or passive) is the same irrespective of the true hypothesis.

The likelihood ratio can be further simplified as

$$L = \mathbb{E}_\pi \exp\left(\sum_{i=1}^m \frac{2y_i \mathrm{tr}(AX_i) - \mathrm{tr}(AX_i)^2}{2\sigma^2}\right).$$

The average risk of the likelihood ratio test

$$R_\pi(T) = 1 - \frac{1}{2}||\mathbb{E}_\pi \mathbb{P}_A - \mathbb{P}_0||_{TV}$$

is determined by the total variation distance between the mixture of alternatives from the null.

By Pinkser's inequality Tsybakov (2009),

$$||\mathbb{E}_\pi \mathbb{P}_A - \mathbb{P}_0||_{TV} \le \sqrt{KL(\mathbb{P}_0, \mathbb{E}_\pi \mathbb{P}_A)/2}$$

and

$$\begin{aligned} KL(\mathbb{P}_0, \mathbb{E}_\pi \mathbb{P}_A) &= -\mathbb{E}_0 \log L \\ &\le -\mathbb{E}_\pi \sum_{i=1}^m \mathbb{E}_0 \frac{2y_i \mathrm{tr}(AX_i) - \mathrm{tr}(AX_i)^2}{2\sigma^2} \\ &= \mathbb{E}_\pi \sum_{i=1}^m \mathbb{E}_0 \frac{\mathrm{tr}(AX_i)^2}{2\sigma^2} \\ &\le \frac{m}{2\sigma^2}||C||_{op}, \end{aligned}$$

where the first inequality follows by applying the Jensen's inequality followed by Fubini's theorem, and the second inequality follows using the fact that $||X_i||_F^2 = 1$, where $C \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$.

To describe the entries of C, consider the invertible map τ from a linear index in $\{1, \dots, n_1 n_2\}$ to an entry of A. Now, $C_{ii} = \mu^2 \mathbb{E}_\pi P_A[A_{\tau(i)} = 1]$ and $C_{ij} = \mu^2 \mathbb{E}_\pi P_A[A_{\tau(i)} = 1, A_{\tau(j)} = 1]$.

To bound the operator norm of C we make two observations. Firstly, because of the contiguous structure of the activation pattern, in any row of C there are at most $k_1 k_2$ non-zero entries. Secondly, each non-zero entry in C is of magnitude at most $\mu^2 k_1 k_2/(n_1 - k_1)(n_2 - k_2)$.

Now, noting that

$$||C||_{op} \leq \max_j \sum_k |C_{jk}| \leq \mu^2 k_1^2 k_2^2/(n_1 - k_1)(n_2 - k_2)$$

from which we obtain a bound on the KL divergence.

Now, this gives us that

$$R_\pi(T) \geq 1 - k_1 k_2 \mu \sqrt{\frac{m}{16(n_1 - k_1)(n_2 - k_2)}}$$

proving the lower bound on the minimax risk.

A.2. Proof of Theorem 2. Define $t = \frac{1}{\sqrt{m}} \sum_{i=1}^m y_i$. It is easy to see that under H_0, $t \sim \mathcal{N}(0, \sigma^2)$ while under H_1, $t \sim \mathcal{N}(\sqrt{\frac{m}{n_1 n_2}} k_1 k_2 \mu, \sigma^2)$. The theorem now follows from an application of standard Gaussian tail bounds.

A.3. Proof of Theorem 3 . Without loss of generality we assume $k_1 \leq k_2$. Consider, two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, where $\mathbb{P}_1$ is induced by matrix A_1 when the activation block $B = B_1 = [1, \ldots, k_1][1, \ldots, k_2]$ and $\mathbb{P}_2$ is induced by matrix A_2 when the activation block $B = B_2 = [1, \ldots, k_1][2, \ldots, k_2 + 1]$.

Following the proof of Theorem 5.

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2) &= \mathbb{E}_{\mathbb{P}_1} \log \frac{\mathbb{P}_1}{\mathbb{P}_2} \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^m (\mathrm{tr}(A_2 X_i) - \mathrm{tr}(A_1 X_i))^2 \\ &= \frac{\mu^2}{\sigma^2} \frac{m k_1}{n_1 n_2} \end{aligned}$$

using the fact that X_i is a random Gaussian matrix with independent entries of variance $\frac{1}{n_1 n_2}$.

Now, noting that the minimax risk

$$R \geq 1 - \sqrt{\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_2)/8}$$

For the second part of the theorem, we consider $\mathbb{P}_2, \ldots, \mathbb{P}_{t+1}$, where $t = (n_1 - k_1)(n_2 - k_2)$, each of which is induced by a B which does not overlap with B_1.

The same calculation now gives

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_j) \le \frac{\mu^2}{\sigma^2}\frac{mk_1k_2}{n_1n_2}$$

Now, applying the multiple hypothesis version of Fano's inequality (see Theorem 2.5 in Tsybakov, 2009) we arrive at the second part of the theorem.

A.4. Proof of Theorem 4. Let $z_{i,B} = \sum_{(a,b)\in B} x_{i,ab}$ and $\mathbf{z}_B = (z_{1,B}, \dots, z_{m,B})'$. With this, we can write the loss function defined in Eq. (4.1) as

$$f(B) := \min_{\mu} ||\mu \mathbf{z}_B - \mathbf{y}||_2^2. \tag{A.1}$$

Let $\Delta(B) = f(B) - f(B^*)$ and observe that an error is made if $\Delta(B) < 0$ for $B \neq B^*$. Therefore,

$$\mathbb{P}[\text{error}] = \mathbb{P}[\cup_{B\in\mathcal{B}\backslash B^*}\{\Delta(B) < 0\}].$$

Under the conditions of the theorem, we will show that $\Delta(B) > 0$ for all $B \in \mathcal{B}\backslash B^*$ with large probability.

The following lemma shows that for any fixed B, the event $\{\Delta(B) < 0\}$ occurs with exponentially small probability.

LEMMA 7. *Fix any $B \in \mathcal{B}\backslash B^*$. Then*

$$\mathbb{P}[\Delta(B) < 0] \le \exp\left(-c_1\frac{(\mu^*)^2 m|B^*\backslash B|}{\sigma^2 n_1 n_2}\right) + c_2\exp(-c_3 m). \tag{A.2}$$

Note that, under the assumptions of the theorem, the first term in Eq. (A.2) dominates the second term and hence will be put into the constant c_1.

Define $N(l) = |\{B \in \mathcal{B} : |B\Delta B*| = l\}|$ to be the number of elements in $\mathcal{B}$ whose with symmetric difference with B^* is equal to l. Note that $N(l) = \mathcal{O}(1)$ for any l. Using the union bound

(A.3)

$$\begin{aligned}
&\mathbb{P}[\cup_{B\in\mathcal{B}}\{\Delta(B) < 0\}] \\
&\le \sum_{B\in\mathcal{B},|B\Delta B^*|=2k_1k_2} \exp\left(-c_1\frac{\mu^2 k_1k_2 m}{\sigma^2 n_1 n_2}\right) + \sum_{l<2k_1k_2} N(l)\exp\left(-c_1\frac{\mu^2 l m}{\sigma^2 n_1 n_2}\right) \\
&\le c_2(n_1-k_1)(n_2-k_2)\exp\left(-c_1\frac{\mu^2 k_1k_2 m}{\sigma^2 n_1 n_2}\right) + c_3k_1k_2\exp\left(-c_1\frac{\mu^2\min(k_1,k_2)m}{\sigma^2 n_1 n_2}\right).
\end{aligned}$$

Choosing

$$\mu = c_1\sigma\sqrt{\frac{n_1n_2}{m}\log(2/\delta)\max\left(\frac{\log\max(k_1,k_2)}{\min(k_1,k_2)}, \frac{\log\max(n_1-k_1,n_2-k_2)}{k_1k_2}\right)}$$

each term in Eq. (A.3) will be smaller than $\delta/2$, with an appropriately chosen constant c_1.

We finish the proof of the theorem, by proving Lemma 7.

PROOF OF LEMMA 7. For any $B \in \mathcal{B}$, let

$$\begin{aligned}\widehat{\mu}_B &= \underset{\mu}{\operatorname{argmin}}\, ||\mu \mathbf{z}_B - \mathbf{y}||_2^2 \\ &= ||\mathbf{Z}_B||_2^{-2}\mathbf{Z}_B'\mathbf{Y}.\end{aligned}$$

Note that $\widehat{\mu}_{B^*} = \mu + ||\mathbf{Z}_B||_2^{-2}\mathbf{Z}_B'\boldsymbol{\epsilon}$.

Let

$$\begin{aligned}\mathbf{H}_B &= ||\mathbf{Z}_B||_2^{-2}\mathbf{Z}_B\mathbf{Z}_B' \\ \mathbf{H}_B^{\perp} &= \mathbf{I} - ||\mathbf{Z}_B||_2^{-2}\mathbf{Z}_B\mathbf{Z}_B'\end{aligned}$$

be the projection matrices and write

$$\begin{aligned}f(B^*) &= ||\mathbf{H}_{B^*}^{\perp}\epsilon||_2^2 \\ f(B) &= ||\mathbf{H}_B^{\perp}(\mathbf{Z}_{B^*}\mu^* + \boldsymbol{\epsilon})||_2^2 = ||\mathbf{H}_B^{\perp}\epsilon||_2^2 + (\mu^*)^2||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}||_2^2 + 2\boldsymbol{\epsilon}'\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}\mu^*.\end{aligned}$$

Now,

$$\boldsymbol{\Delta}(B) = \underbrace{||\mathbf{H}_B^{\perp}\epsilon||_2^2 - ||\mathbf{H}_{B^*}^{\perp}\epsilon||_2^2}_{T_1} + \underbrace{(\mu^*)^2||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}||_2^2 + 2\boldsymbol{\epsilon}'\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}\mu^*}_{T_2}.$$

Let $V_1, V_2 \sim \chi^2_{m-1}$. Observe that $T_1 \sim \sigma^2(V_1 - V_2)$.

(A.4)

$$\mathbb{P}\left[|T_1| \geq \frac{\sigma^2(m-1)\epsilon}{2}\right] \leq 2\mathbb{P}\left[|\chi^2_{m-1} - m + 1| \geq \frac{(m-1)\epsilon}{4}\right] \leq 2\exp\left(-\frac{3(m-1)\epsilon^2}{256}\right)$$

using Eq. (B.4), as long as $\epsilon \in [0, 2)$.

To analyze the term T_2, we condition on $\mathbf{X}$, so that

$$T_2|\mathbf{X} \sim \mathcal{N}(\widetilde{\mu}, 4\sigma^2\widetilde{\mu})$$

where $\widetilde{\mu} = (\mu^*)^2||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}||_2^2$. This gives

$$\mathbb{P}[T_2 \leq \widetilde{\mu}/2|\mathbf{X}] = \mathbb{P}[\mathcal{N}(0,1) \geq \sqrt{\widetilde{\mu}}/(4\sigma)|\mathbf{X}].$$

Next, we show how to control $||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}||_2^2$. Writing $\mathbf{Z}_{B^*} = \mathbf{Z}_B - \mathbf{Z}_{B\backslash B^*} + \mathbf{Z}_{B^*\backslash B}$, simple algebra gives

$$
\begin{aligned}
&||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}||_2^2 \\
&= ||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*\backslash B}||_2^2 + ||\mathbf{H}_B^{\perp}\mathbf{Z}_{B\backslash B^*}||_2^2 - 2\mathbf{Z}'_{B^*\backslash B}\mathbf{H}_B^{\perp}\mathbf{Z}_{B\backslash B^*} \\
&= ||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*\backslash B}||_2^2 + ||\mathbf{Z}_{B\backslash B^*} - \mathbf{Z}_{B^*\backslash B}||_2^2 - ||\mathbf{Z}_{B^*\backslash B}||_2^2 - \frac{((\mathbf{Z}_{B\backslash B^*} - \mathbf{Z}_{B^*\backslash B})'\mathbf{Z}_B)^2 - (\mathbf{Z}'_{B^*\backslash B}\mathbf{Z}_B)^2}{||\mathbf{Z}_B||_2^2} \\
&\geq ||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*\backslash B}||_2^2 + ||\mathbf{Z}_{B\backslash B^*} - \mathbf{Z}_{B^*\backslash B}||_2^2 - ||\mathbf{Z}_{B^*\backslash B}||_2^2 - \frac{((\mathbf{Z}_{B\backslash B^*} - \mathbf{Z}_{B^*\backslash B})'\mathbf{Z}_B)^2}{||\mathbf{Z}_B||_2^2}.
\end{aligned}
$$

Define the event

$$
\begin{aligned}
\mathcal{E}(\epsilon) = &\left\{||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*\backslash B}||_2^2 \geq \frac{(1-\epsilon)(m-1)|B^*\backslash B|}{n_1n_2}\right\} \bigcap \left\{||\mathbf{Z}_{B\backslash B^*} - \mathbf{Z}_{B^*\backslash B}||_2^2 \geq \frac{(1-\epsilon)2m|B^*\backslash B|}{n_1n_2}\right\} \\
&\bigcap \left\{||\mathbf{Z}_{B^*\backslash B}||_2^2 \leq \frac{(1+\epsilon)m|B^*\backslash B|}{n_1n_2}\right\} \bigcap \left\{||\mathbf{Z}_B||_2^2 \geq \frac{(1-\epsilon)m|B|}{n_1n_2}\right\} \\
&\bigcap \left\{|(\mathbf{Z}_{B\backslash B^*} - \mathbf{Z}_{B^*\backslash B})'\mathbf{Z}_B| \leq \frac{(1+\epsilon)m|B^*\backslash B|}{n_1n_2}\right\},
\end{aligned}
$$

such that, using the concentration results in Appendix B,

$$
\mathbb{P}[\mathcal{E}(\epsilon)^C] \leq c_1 \exp(-c_2 m \epsilon^2).
$$

On the event $\mathcal{E}(\epsilon)$ we have that

$$
\begin{aligned}
||\mathbf{H}_B^{\perp}\mathbf{Z}_{B^*}||_2^2 &\geq \frac{m|B^*\backslash B|}{n_1n_2}\left[3(1-\epsilon) - (1+\epsilon) - \frac{(1+\epsilon)^2}{1-\epsilon}\frac{|B^*\backslash B|}{|B|}\right] - \frac{(1-\epsilon)|B^*\backslash B|}{n_1n_2} \\
&\geq c_1\frac{m|B^*\backslash B|}{n_1n_2}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}[T_2 \leq \widetilde{\mu}/2|\mathbf{X}] &\leq \mathbb{P}\left[\mathcal{N}(0,1) \geq c_1\frac{\mu^*}{\sigma}\sqrt{\frac{m|B^*\backslash B|}{n_1n_2}}\right] + \mathbb{P}[\mathcal{E}^C] \\
&\leq \exp\left(-c_1\frac{(\mu^*)^2 m|B^*\backslash B|}{\sigma^2 n_1n_2}\right) + c_2\exp(-c_3 m\epsilon^2).
\end{aligned} \tag{A.5}
$$

Combining Eq. (A.4) and Eq. (A.5) completes the proof.

□

A.5. Proof of Theorem 5. The proof will proceed via two separate constructions. At a high level these constructions are intended to capture the difficulty of exactly and approximately localizing the activation block.

Construction 1 - approximate localization: Let us define three distributions: $\mathbb{P}_0$ corresponding to no bicluster, $\mathbb{P}_1$ which is a uniform mixture over the distributions induced by having the top-left corner of the bicluster in the left half of the matrix and $\mathbb{P}_2$ which is a uniform mixture over the distributions induced by having the top-left corner of the bicluster in the right half of the matrix.

We first upper bound the total variation between $\mathbb{P}_1$ and $\mathbb{P}_2$. This results directly in a lower bound for the problem of distinguishing whether the top-left corner of the bicluster is in the left or right half of the matrix, which in turn is a lower bound for the localization of the bicluster.

Now notice that,

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

Notice that $KL(\mathbb{P}_0, \mathbb{P}_1)$ is exactly the quantity we have to upper bound to produce a lower bound on the signal strength for detecting whether there is a bicluster in the left half of the matrix or not. At least from a lower bound perspective this reduces the problem of localization to that of detection. We can now apply a slight modification of the proof of Theorem 1 to obtain that

$$KL(\mathbb{P}_0, \mathbb{P}_1) = KL(\mathbb{P}_0, \mathbb{P}_2) \leq \frac{m\mu^2 k_1^2 k_2^2}{(n_1 - k_1)(n_2/2 - k_2)}$$

Noting that the minimax risk R for distinguishing $\mathbb{P}_1$ from $\mathbb{P}_2$

$$R = 1 - \frac{1}{2}||\mathbb{P}_1 - \mathbb{P}_2||_{TV} \geq 1 - \sqrt{\frac{m\mu^2 k_1^2 k_2^2}{2(n_1 - k_1)(n_2/2 - k_2)}}$$

Construction 2 - exact localization: Without loss of generality we assume $k_1 \leq k_2$. Consider, two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, where $\mathbb{P}_1$ is induced by matrix A_1 when the activation block $B = B_1 = [1, \ldots, k_1][1, \ldots, k_2]$ and $\mathbb{P}_2$ is induced by matrix A_2 when the activation block $B = B_2 = [1, \ldots, k_1][2, \ldots, k_2 + 1]$.

Now, following the same argument as in the proof of Theorem 1, we have

$$\begin{aligned} KL(\mathbb{P}_1, \mathbb{P}_2) &= \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^{m} \left(-\frac{1}{2\sigma^2} \left[(y_i - \mathrm{tr}(A_1 X_i))^2 - (y_i - \mathrm{tr}(A_2 X_i))^2 \right] \right) \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^{m} \left[\mathrm{tr}(A_2 X_i)^2 - \mathrm{tr}(A_1 X_i)^2 + 2 y_i \mathrm{tr}(A_1 X_i) - 2 y_i \mathrm{tr}(A_2 X_i) \right] \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^{m} \Big(\underbrace{\mathrm{tr}(A_2 X_i) - \mathrm{tr}(A_1 X_i)}_{t_i} \Big)^2 = \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_1} \sum_{i=1}^{m} t_i^2 \end{aligned}$$

Now, with some abuse of notation,

$$\begin{aligned} t_i &= \mu \left(\sum_{j \in B_1 \setminus B_2} X_{ij} - \sum_{j \in B_2 \setminus B_1} X_{ij} \right) \\ &\leq \mu \left(\sum_{j \in B_1 \Delta B_2} |X_{ij}| \right) \end{aligned}$$

By using Cauchy-Schwarz we get

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

since $||X_i||_F^2 = 1$.

This gives us that,

$$KL(\mathbb{P}_1, \mathbb{P}_2) \leq \frac{m k_1 \mu^2}{\sigma^2}$$

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

$$KL(\mathbb{P}_1, \mathbb{P}_2) \leq \frac{m \min(k_1, k_2) \mu^2}{\sigma^2}$$

Once again noting (by Pinsker's theorem),

$$R \geq 1 - \sqrt{KL(\mathbb{P}_1, \mathbb{P}_2)/8} \geq 1 - \sqrt{\frac{m \min(k_1, k_2) \mu^2}{8\sigma^2}}$$

Combining the approximate and exact localization bounds we get,

$$R \geq \max\left(1 - \sqrt{\frac{m\min(k_1,k_2)\mu^2}{8\sigma^2}}, 1 - \sqrt{\frac{m\mu^2 k_1^2 k_2^2}{2(n_1-k_1)(n_2/2-k_2)}}\right)$$

Thus, we get for any $0 < \alpha < 1$, $R \geq \alpha$ if

$$\min\left(\sqrt{\frac{m\min(k_1,k_2)\mu^2}{8\sigma^2}}, \sqrt{\frac{m\mu^2 k_1^2 k_2^2}{2(n_1-k_1)(n_2/2-k_2)}}\right) \leq 1 - \alpha$$

A.6. Proof of Theorem 6. As with the lower bound the localization algorithm and analysis is naturally divided into two phases. An approximate localization phase and an exact localization one. We will analyze each of these in turn. To ease presentation we will assume n_1 is a dyadic multiple of $2k_1$ and n_2 a dyadic multiple of $2k_2$. Straightforward modifications are possible when this is not the case.

Approximate localization: The approximate localization phase proceeds by a modification of the compressive binary search (CBS) procedure of Malloy and Nowak (2012) (see also Davenport and Arias-Castro (2012)) on the matrix A.

We will run this modified CBS procedure *twice* on two sets of blocks of the matrix A. The first set consists of the blocks

$$\mathcal{D}_1 \equiv \{B_{11} = [1,\ldots,2k_1][1,\ldots,2k_2] \cup B_{12} = [2k_1+1,\ldots,4k_1][1,\ldots,2k_2] \cup \ldots \\ \ldots \cup B_{1n_1n_2/4k_1k_2} = [n_1 - 2k_1,\ldots,n_1][n_2 - 2k_2,\ldots,n_2]\big\}$$

The second set consists of the blocks

$$\mathcal{D}_2 \equiv \{B_{21} = [k_1,\ldots,3k_1][k_2,\ldots,3k_2] \cup B_{22} = [3k_1+1,\ldots,5k_1][k_2,\ldots,3k_2] \cup \ldots \\ \cup \ldots B_{2n_1n_2/4k_1k_2} = [n1-k1,...,n1,1,\ldots,k_1][n2-k2,...,n2,1,\ldots,k_2]\big\}$$

Notice that the entire block of activation is always *fully* contained in one of these blocks. The output of the CBS procedure when run on these two collections is two blocks - one from the first collection and the second from the second collection. We define an approximate localization *error* to be the event in which neither of the two blocks returned fully contains the block of activation.

Without loss of generality let us assume that the activation block is fully contained in some block from the first collection. Once we have fixed the collection of blocks the CBS procedure is invariant to reordering of the blocks, so without loss of generality we can consider the case when the activation block is contained in B_{11}.

The analysis proceeds exactly as in Malloy and Nowak (2012), we detail the differences arising from having a block of activation as opposed to a single activation in a vector. Notice, that the binary search procedure on the first collection of blocks proceeds for

$$s_0 \equiv \log\left(\frac{n_1 n_2}{4k_1 k_2}\right)$$

rounds. Now, we can bound the probability of error of the procedure by a union bound as

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

where

$$w^s \sim \mathcal{N}\left(\frac{m_s 2^{(s-1)/2} k_1 k_2 \mu}{\sqrt{n_1 n_2}}, m_s \sigma^2\right)$$

Recall, the allocation scheme: for $m \geq 2s_0$, $m_s \equiv \lfloor (m - s_0) s 2^{-s-1} \rfloor + 1$ and observe that $\sum_{s=1}^{s_0} m_s \leq m$

Now, using the Gaussian tail bound

$$P[N(0,1) > t] \leq \frac{1}{2}\exp(-t^2/2)$$

we see that

$$\mathbb{P}_e \leq \frac{1}{2}\sum_{s=1}^{s_0} \exp\left(-\frac{m_s 2^s k_1^2 k_2^2 \mu^2}{4 n_1 n_2 \sigma^2}\right)$$

Now, observe that $m_s \geq (m - s_0) s 2^{-s-1}$ and $m \geq 2s_0$, so $m_s \geq m s 2^{-s-2}$. It is now straightforward to verify that if

$$\mu \geq \sqrt{\frac{16\sigma^2 n_1 n_2}{m k_1^2 k_2^2} \log\left(\frac{1}{2\delta} + 1\right)}$$

we have $\mathbb{P}_e \leq \delta$.

Let us revisit what we have shown so far: if μ is large enough then one of the two runs of the CBS procedure will return a block of size $(2k_1 \times 2k_2)$ which fully contains the block of activation, with probability at least $1 - 2\delta$.

Exact localization: In the $1 - \delta$ probability event described above, we have a block of at most $(4k_1 \times 4k_2)$ which contains the full block of activation (for simplicity we disregard the fact that we know that the block is actually in one of two $(2k_1 \times 2k_2)$ blocks).

Let us first identify the active columns. First, notice that one of the first, k_2+1st, $2k_2+1$st or $3k_2+1$st column must be active. Let us devote $4m$ measurements to identifying the active column amongst these. The procedure is straightforward: measure each column m times, and pick the largest.

It is easy to show that the active column results in a draw from $\mathcal{N}(\sqrt{k_1}\mu m/2, m\sigma^2)$ and the non-active columns result in draws from $\mathcal{N}(0, m\sigma^2)$.

Using the same Gaussian tail bound as before it is easy to show that if

$$\mu \geq \sqrt{\frac{32\sigma^2}{k_1 m}\log(2/\delta)}$$

we successfully find the active column with probability at least $1-\delta$.

So far, we have identified an active column and localized the columns of the activation block to one of $2k_2$ columns. We will use m more measurements to find the remaining active 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 $m/(3\log k_2)$ measurements to each column. We will need to threshold these measurements at

$$\sqrt{\log\left(\frac{3\log k_2}{\delta}\right)\frac{2m\sigma^2}{3\log k_2}}$$

and declare a row as active if its average is larger than this.

It is easy to show that this binary search procedure successfully finds all active columns with probability at least $1-\delta$ if

$$\mu \geq \sqrt{\frac{32\sigma^2 \log k_2}{mk_1}\log\left(\frac{3\log k_2}{\delta}\right)}$$

We repeat this procedure to identify the active rows.

Putting everything together: Total number of measurements used:

1. Two rounds of CBS: $2m$
2. Identifying first active column and first active row: $8m$
3. Identifying remaining active rows and columns: $2m$

This is a total of $12m$ measurements. Each of these steps fails with a probability at most δ, for a total of 6δ.

Now, re-adjusting constants we obtain, if

$$\mu \geq \max\left(\sqrt{\frac{192\sigma^2 n_1 n_2}{m k_1^2 k_2^2}\log\left(\frac{3}{\delta}+1\right)}, \sqrt{\frac{384\sigma^2 \log\max(k_1,k_2)}{m\min(k_1,k_2)}\log\left(\frac{18\log\max(k_1,k_2)}{\delta}\right)}\right)$$

then we successfully localize the matrix with probability at least $1-\delta$.

Stated more succinctly we require

$$\mu \geq \widetilde{O}\left(\max\left(\sqrt{\frac{\sigma^2 n_1 n_2}{m k_1^2 k_2^2}}, \sqrt{\frac{\sigma^2}{\min(k_1,k_2)m}}\right)\right).$$

This matches the lower bound up to $\log k$ factors.

APPENDIX B: COLLECTION OF CONCENTRATION RESULTS

In this section, we collect useful results on tail bounds of various random quantities used throughout the paper. We start by stating a lower and upper bound on the survival function of the standard normal random variable. Let $Z \sim \mathcal{N}(0,1)$ be a standard normal random variable. Then for $t > 0$

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

B.1. Tail bounds for Chi-squared variables. Throughout the paper we will often use one of the following tail bounds for central χ^2 random variables. These are well known and proofs can be found in the original papers.

LEMMA 8 (Laurent and Massart (2000)). *Let $X \sim \chi_d^2$. For all $x \geq 0$,*

$$\mathbb{P}[X - d \geq 2\sqrt{dx} + 2x] \leq \exp(-x) \tag{B.2}$$

$$\mathbb{P}[X - d \leq -2\sqrt{dx}] \leq \exp(-x). \tag{B.3}$$

LEMMA 9 (Johnstone and Lu (2009)). *Let $X \sim \chi_d^2$, then*

$$\mathbb{P}[|d^{-1}X - 1| \geq x] \leq \exp(-\frac{3}{16}dx^2), \quad x \in [0, \frac{1}{2}). \tag{B.4}$$

The following result provide a tail bound for non-central χ^2 random variable with non-centrality parameter ν.

LEMMA 10 (Birgé (2001)). *Let $X \sim \chi^2_d(\nu)$, then for all $x > 0$*

$$\mathbb{P}[X \geq (d+\nu) + 2\sqrt{(d+2\nu)x} + 2x] \leq \exp(-x) \tag{B.5}$$

$$\mathbb{P}[X \leq (d+\nu) - 2\sqrt{(d+2\nu)x}] \leq \exp(-x). \tag{B.6}$$

Using the above results, we have a tail bound for sum of product-normal random variables.

LEMMA 11. *Let $Z = (Z_a, Z_b) \sim \mathcal{N}_2(0, 0, \sigma_{aa}, \sigma_{bb}, \sigma_{ab})$ be a bivariate Normal random variable and let $(z_{ia}, z_{ib}) \overset{iid}{\sim} Z$, $i = 1, \dots, n$. Then or all $t \in [0, \nu_{ab}/2)$*

$$\mathbb{P}\left[\left|n^{-1}\sum_i z_{ia}z_{ib} - \sigma_{ab}\right| \geq t\right] \leq 4\exp\left(-\frac{3nt^2}{16\nu_{ab}^2}\right), \tag{B.7}$$

where $\nu_{ab} = \max\{(1-\rho_{ab})\sqrt{\sigma_{aa}\sigma_{bb}}, (1+\rho_{ab})\sqrt{\sigma_{aa}\sigma_{bb}}\}$.

PROOF. Let $z'_{ia} = z_{ia}/\sqrt{\sigma_{aa}}$. Then using (B.4)

$$\begin{aligned}
&\mathbb{P}[|\frac{1}{n}\sum_{i=1}^n z_{ia}z_{ib} - \sigma_{ab}| \geq t] \\
&\quad= \mathbb{P}[|\frac{1}{n}\sum_{i=1}^n z'_{ia}z'_{ib} - \rho_{ab}| \geq \frac{t}{\sqrt{\sigma_{aa}\sigma_{bb}}}] \\
&\quad= \mathbb{P}[|\sum_{i=1}^n((z'_{ia}+z'_{ib})^2 - 2(1+\rho_{ab})) - ((z'_{ia}-z'_{ib})^2 - 2(1-\rho_{ab}))| \geq \frac{4nt}{\sqrt{\sigma_{aa}\sigma_{bb}}}] \\
&\quad\leq \mathbb{P}[|\sum_{i=1}^n((z'_{ia}+z'_{ib})^2 - 2(1+\rho_{ab}))| \geq \frac{2nt}{\sqrt{\sigma_{aa}\sigma_{bb}}}] \\
&\qquad+ \mathbb{P}[|\sum_{i=1}^n((z'_{ia}-z'_{ib})^2 - 2(1-\rho_{ab}))| \geq \frac{2nt}{\sqrt{\sigma_{aa}\sigma_{bb}}}] \\
&\quad\leq 2\mathbb{P}[|\chi^2_n - n| \geq \frac{nt}{\nu_{ab}}] \leq 4\exp(-\frac{3nt^2}{16\nu_{ab}^2}),
\end{aligned}$$

where $\nu_{ab} = \max\{(1-\rho_{ab})\sqrt{\Sigma_{aa}\Sigma_{bb}}, (1+\rho_{ab})\sqrt{\Sigma_{aa}\Sigma_{bb}}\}$ and $t \in [0, \nu_a/2)$. □

COROLLARY 12. *Let Z_1 and Z_2 be two independent standard Normal random variables and let $X_i \overset{iid}{\sim} Z_1Z_2$, $i = 1 \dots n$. Then for $t \in [0, 1/2)$*

$$\mathbb{P}[|n^{-1}\sum_{i\in[n]} X_i| > t] \leq 4\exp(-\frac{3nt^2}{16}). \tag{B.8}$$

REFERENCES

E. Arias-Castro. Detecting a vector based on linear measurements. *Electronic Journal of Statistics*, 6:547–558, 2012. ISSN 1935-7524.

E. Arias-Castro, E.J. Candès, and M.A. Davenport. On the fundamental limits of adaptive sensing. *arXiv:1111.4646*, 2011.

R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

L. Birgé. An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, 36:113–133, 2001. ISSN 0749-2170.

C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *ArXiv e-prints*, September 2011.

E.J. Candès and M.A. Davenport. How well can we estimate a sparse vector? *arXiv:1104.5246*, 2011.

E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

E.J. Candès and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 21, 2008.

M.A. Davenport and E. Arias-Castro. Compressive binary search. *arXiv:1202.0937*, 2012.

D.L. Donoho. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D.L. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004.

J.D. Haupt, R.G. Baraniuk, R.M. Castro, and R.D. Nowak. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Proceedings of the 43rd Asilomar conference on Signals, systems and computers*, pages 1551 –1555, 2009.

J. Jang, D. Brumley, and S. Venkataraman. Bitshread: feature hashing malware for scalable triage and semantic analysis. In *Proceedings of the 18th ACM conference on Computer and communications security*, CCS '11, pages 309–320, 2011. ISBN 978-1-4503-0948-6. .

I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 909–917, 2011.

V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000. .

M.L. Malloy and R.D. Nowak. Near-optimal compressive binary search. *arXiv:1203.1804*, 2012.

W. C. Moore, D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, Jr. R. D'Agostino, M. Castro, D. Curran-Everett, A. M. Fitzpatrick, B. Gaston, N. N. Jarjour, R. Sorkness, W. J. Calhoun, K. Fan Chung, S. A. A. Comhair, R. A. Dweik, E. Israel, S. P. Peters,

W. W. Busse, S. C. Erzurum, and E. R. Bleecker. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. *Am J Respir Crit Care Med*, 181(4):315–323, February 2010.

S. Negahban and M.J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

A. Soni and J. Haupt. Efficient adaptive compressive sensing using sparse hierarchical learned dictionaries. *arXiv:1111.6923*, 2011.

A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009. ISBN 9780387790510.

M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009a. ISSN 0018-9448. .

M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009b.

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 Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(4):339–354, 2005.

S. Balakrishnan, M. Kolar, A. Rinaldo and A. Singh
Carnegie Mellon University
6105 Gates Hillman Complex
5000 Forbes Ave
Pittsburgh, PA 15217
USA
E-mail: sbalakri@cs.cmu.edu
mladenk@cs.cmu.edu
arinaldo@stat.cmu.edu
aarti@cs.cmu.edu