Detection of correlations with adaptive sensing

Document status and date:
Published: 01/01/2014

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Download date: 04. Oct. 2023

Detection of Correlations With Adaptive Sensing

Rui M. Castro, Gábor Lugosi, and Pierre-André Savalle

Abstract—The problem of detecting correlations from samples of a high-dimensional Gaussian vector has recently received a lot of attention. In most existing work, detection procedures are provided with a full sample. However, following common wisdom in experimental design, the experimenter may have the capacity to make targeted measurements in an on-line and adaptive manner. In this paper, we investigate such adaptive sensing procedures for detecting positive correlations. It is shown that, using the same number of measurements, adaptive procedures are able to detect significantly weaker correlations than their nonadaptive counterparts. We also establish minimax lower bounds that show the limitations of any procedure.

Index Terms—Sequential testing, adaptive sensing, sparse covariance matrices, sparse principal component analysis, high-dimensional detection.

I. INTRODUCTION

IN THIS paper we consider a statistical testing problem related to anomaly detection: the detection of correlations between signals. In the general problem of anomaly detection, one aims to identify unexpected activity in data. It has applications in numerous domains [14], such as finance [9], computer security [21], health monitoring [29], or detection of activity in sensor networks [24], [33], [39]. In many situations, anomalies can be detected by looking at unusual signal values at any of the sensors. For instance, a home security alarm is usually comprised of various infrared or related sensors, and an alert is raised as soon as a single sensor detects an unusual signal. However, in other situations, when signals are "weak", they may never appear anomalous in isolation, and anomalies may only be detected when considering the signals together as a collection. This type of phenomena may be referred to as either contextual anomaly detection [35], or collective anomaly detection [34], depending on the setup. A prototypical example of such a problem is the detection of Distributed

Denial-of-Service (DDoS) attacks in computer networks, which has become an important challenge in recent years [32], [37], [40]. In a DDoS attack, the attacker usually controls a large number of computers distributed around the world. These machines are used to simultaneously send requests to a target server, which is then flooded by the amount of packets, and can become unavailable as a result. As a side effect, this type of attack can produce high volumes of traffic in various parts of the worldwide internet infrastructure. However, packets sent by the attacker through the machines that he/she controls cannot usually be detected as anomalous in isolation [27], and detection of DDoS requires to correlate signals obtained at different points in the network. Collective anomalies also appear, for instance, in the context of detection of the outbreak of diseases [28]. Another important type of anomaly detection problem appears when dealing with sensor data arranged on a two-dimensional grid (e.g., loop detectors in lanes of road networks, or wireless sensor networks [2]). In this case, collective anomalies may be characterised by *neighbouring* signals being correlated. Besides anomaly detection, detection of correlations is also of interest to assess to what extent dimensionality reduction can be performed on a data stream. Reduction of dimensionality is a workhorse of data analysis, and there has been a strong recent interest in modifying principal component analysis to deal with high-dimensional data [10], [12], [26]. Testing when this type of transformation is justified is thus an important problem.

In this work, we consider a simple correlation model: given multiple observations from a Gaussian multivariate distribution we want to test whether the corresponding covariance matrix is diagonal against non-diagonal alternatives. Such problems have recently received a lot of attention in the literature, where different models and choices of non-diagonal covariance alternatives were considered [4], [5], [10], [12], [20]. We consider the detection of sparse *positive* correlations, which has been treated in the case of a unique multivariate sample [4], or of multiple samples [5]. However, this paper deviates from the existing literature in that we consider an *adaptive sensing* or *sequential experimental design* setting. More precisely, data is collected in a sequential and adaptive way, where data collected at earlier stages informs the collection of data in future stages. Adaptive sensing has been studied in the context of other detection and estimation problems, such as in detection of a shift in the mean of a Gaussian vector [13], [19], in compressed sensing [6], [13], [18], in experimental design, optimization with Gaussian processes [36], and in active learning [15]. Adaptive sensing procedures are quite flexible, as the data collection procedure can be "steered" to ensure

R. M. Castro is with the Department of Mathematics, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands (e-mail: rmcastro@tue.nl).

G. Lugosi is with ICREA and Department of Economics, Pompeu Fabra University, Barcelona 08005, Spain (e-mail: gabor.lugosi@upf.edu).

P.-A. Savalle is with the Centre de Mathématiques et de leurs Applications, École Normale Supérieure de Cachan, Cachan 94230, France, and also with the Laboratoire de Mathématiques Appliquées aux Systèmes, École Centrale Paris, Chatenay-Malabry 92290, France (e-mail: pierre-andre.savalle@ecp.fr).

most collected data provides important information. As a consequence, procedures based on adaptive sensing are often associated with better detection or estimation performances than those based on non-adaptive sensing with a similar measurement budget. In this paper, our objective is to determine whether this is also the case for detection of sparse positive correlations, and if so, to quantify how much can be gained.

A. Model

Let $U^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$ be independent and identically distributed (i.i.d.) normal random vectors with zero mean and covariance matrix Σ_S, where S is a subset of $[n] = \{1, \ldots, n\}$. Let $\rho > 0$ and define the covariance matrix as

$$(\Sigma_S)_{i,j} = \begin{cases} 1, & i = j \\ \rho, & i \neq j, \text{ with } i, j \in S \\ 0, & \text{otherwise.} \end{cases}$$

Our main goal is to solve the hypothesis testing problem

$$H_0 : S = \emptyset$$
$$H_1 : S \in \mathcal{C},$$

where $\mathcal{C}$ is some class of non-empty subsets of $\{1, \ldots, n\}$, each of size k. In other words, under the alternative hypothesis, there exists an unknown subset $S \in \mathcal{C}$ such that corresponding components are positively correlated with strength $\rho > 0$. We often refer to the elements of S as the subset of *contaminated* coordinates. The model of correlations we consider appears naturally in the problem of detecting a sparse signal embedded in noise. Indeed, with (Y_i^t) and N^t being independent standard normal random variables, and

$$U_i^t = \begin{cases} Y_i^t, & i \notin S \\ \sqrt{1-\rho} Y_i^t + \sqrt{\rho} N^t, & i \in S \end{cases}$$

for some $S \in \mathcal{C}$, then the vectors U^t are independent multi-variate zero-mean normal vectors with covariance matrix Σ_S. The variable N_t represent a common signal present at each contaminated coordinate and Y_i^t the additive white noise. In all cases we assume that the cardinality of each $S \in \mathcal{C}$ is the same: $|S| = k$. We consider the following types of classes $\mathcal{C}$ for the contaminated coordinates:

- **k-intervals**: all sets of k contiguous coordinates, of the form $\{z, z+1, \ldots, z+k-1\}$ for some $1 \leq z \leq n-k+1$; this class has size linear in n, and we denote it by $\mathcal{C}_{[k]}$.
- **disjoint k-intervals**: the class $\mathcal{D}_{[k]}$ defined as

$$\mathcal{D}_{[k]} = \{I_1, \ldots, I_{\lfloor n/k \rfloor}\}$$

where $I_j = \{(j-1)k+1, \ldots, jk\}$, $j \in \{1, \ldots, \lfloor n/k \rfloor\}$.
- **k-sets**: all subsets of $\{1, \ldots, n\}$ of cardinality k. We denote this class by $\mathcal{C}_k$.

In addition, it is of interest for applications to consider settings where the coordinates $\{1, \ldots, n\}$ are laid out according to a two-dimensional grid $[n_1] \times [n_2]$ with $n_1 n_2 = n$, similarly to a spatially arranged array of sensors. Although k-sets still make sense in this setting, the contaminated set can be further assumed in this case to be connected and spatially localized in some sense. The following example is most intuitive:

- **(k_1, k_2)-rectangles**: for $k_1 k_2 = k$, this comprises all sets of the form

$$\{i_0, \ldots, i_0 + k_1 - 1\} \times \{j_0, \ldots, j_0 + k_2 - 1\}$$

for $i_0 \in [n_1 - k_1 + 1]$, $j_0 \in [n_2 - k_2 + 1]$.

Results for rectangles or similar two-dimensional shapes can be obtained easily from our results for k-intervals, and are identical up to constants. We omit the rather straightforward details here.

For any $t = 1, 2, \ldots$ denote by $\mathbb{P}_\emptyset$ the distribution of U^t under the null, and by $\mathbb{P}_S$ the distribution under the alternative with contaminated set $S \in \mathcal{C}$. In addition, for a positive integer q, we denote by $\mathbb{P}^{\otimes q}$ the product measure $\mathbb{P} \otimes \ldots \otimes \mathbb{P}$ with q factors. As previously, we let $[q] = \{1, \ldots, q\}$.

B. Adaptive vs. Non-Adaptive Sensing and Testing

Clearly, the above hypothesis testing problem would be trivial if one has access to an infinite number of i.i.d. samples $(U^t)_{t \in \{1, \ldots, \infty\}}$. Therefore, one must include some further restrictions on the data that is made available for testing. In particular, we only consider testing procedures that make use of at most M entries of the matrix $(U_i^t)_{t \in \{1, \ldots, \infty\}, i \in [n]}$. It is useful to regard this as a matrix with n columns and an infinite number of rows.

The key idea of adaptive sensing is that information gleaned from previous observations can be used to guide the collection of future observations. To formalize this idea consider the following notation: for any subset $A \subseteq [n]$ we denote by $|A|$ the cardinality of A. When A is nonempty we write $U_A = (U_i)_{i \in A} \in \mathbb{R}^{|A|}$ for the subvector of a vector $U \in \mathbb{R}^n$ indexed by coordinates in A. Finally, if U is a random variable taking values in $\mathbb{R}^n$ denote by $\mathbb{P}|_A$ the distribution of U_A.

Let $S \in \mathcal{C} \cup \{\emptyset\}$ be the set of contaminated coordinates, and $M \geq 2$ be an integer. In our model we are allowed to collect information as follows. We consider successive rounds. At round $t \in \mathbb{N}$, one chooses a non-empty *query* subset $A^t \subseteq [n]$ of the components, and observes $U_{A^t}^t$. To avoid technical difficulties later on, we define the observation made at time t as X^t, so that $X_{A^t}^t = U_{A^t}^t$ and $X_{[n] \setminus A^t}^t = \mathbf{0}$. In words, one observes the A^t coordinates of U^t, while the remaining coordinates are completely uninformative. Each successive round proceeds in the same fashion, under the requirement that the budget constraint

$$\sum_{t=1}^{\infty} |A^t| \leq M \tag{1}$$

is satisfied. Note that clearly, the number of rounds is not larger than M. Again, to avoid technical difficulties we assume the total number of rounds to be M in what follows, even if this means $A^t = \emptyset$ for some values of t. See Figure 1 for an illustration.

In our setting, one can select the query sequence randomly and sequentially, and hence, we write the query sequence $(a^1, \ldots, a^M)$ as a realization of a sequence $(A^1, \ldots, A^M)$ of M random subsets of $[n]$, some of which may be empty, and such that $\sum_{t=1}^{M} |A^t| \leq M$.
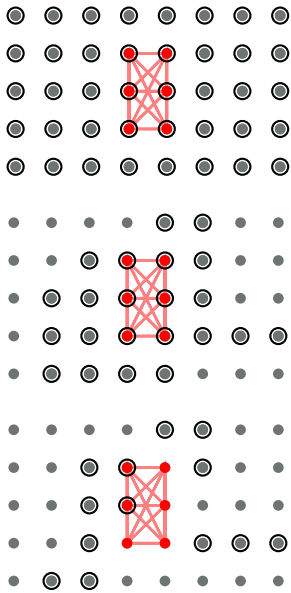
Fig. 1. Adaptive sensing over a two dimensional grid of sensors. The figure illustrates how information can be obtained within the sensing model for $n = 40$ and $k = 6$, under the alternative hypothesis with S being a $(2, 3)$-rectangle in a 8×5 grid. The correlated coordinates form a clique in the graph of correlations, and this is shown through light edges. At every step, the experimenter selects coordinates to be sensed, and these are shown circled. At the first step, the experimenter samples all the coordinates, while at the two subsequent steps, the experimenter reduced the amount of coordinates sampled. This corresponds to a total budget of $|A^1| + |A^2| + |A^3| = 40 + 22 + 13 = 75$ coordinate measurements.

A key aspect of adaptive sensing is that the query at round T may depend on all the information available up to that point. We assume A^t can depend on the history at time $t - 1$, which we denote by $H^{t-1} = (A^j, X^j)_{j \in [t-1]}$. More precisely, we assume A^t is a measurable function of H^{t-1}, and possibly of additional randomization. We call the collection of all the conditional distributions of A^t given H^{t-1} for $t \in [M]$ the *sensing strategy*. In particular, if there is no additional randomization, A^t is a deterministic function of H^{t-1}. We denote the set of all possible adaptive sensing strategies with sensing budget M as $\mathbf{AS}(M)$.

At this point, it is important to formally clarify what is meant by *non-adaptive sensing*. This is simply the scenario where $(A^t)_{t \in [M]}$ is independent of $(U_i^t)_{t \in [M]}, i \in [n]$. In other words, all the decisions regarding the collection of data must be taken before any observations are made. The collection $(A^t)_{t \in [M]}$ is known as a *non-adaptive sensing strategy*. A natural and important choice is *uniform sensing*, where $A^t = [n]$ for $t = 1, \ldots, M/n$ (assume M is divisible by n). In words, one collects $m = M/n$ i.i.d. samples from $\mathbb{P}_S$. This problem has been thoroughly studied in [4]; we summarize some of the main results of [4] in Section I-C.

Now that we have formalized how data is collected, we can perform statistical tests. Formally, a *test* is a measurable binary function $\phi : H^M \mapsto \phi(H^M) \in \{0, 1\}$, that is, a binary function of all the information obtained by the (adaptive or non-adaptive) sensing strategy. The result of the test is $\phi(H^M)$, and if this is one we declare the rejection of the null hypothesis. Finally, an *adaptive testing procedure* is a pair $(\mathcal{A}, \phi)$ where $\mathcal{A}$ is a sensing strategy and ϕ is a test.

For any sensing strategy $\mathcal{A}$ and $S \in \mathcal{C}$, define $\mathbb{P}_{\emptyset}^{\mathcal{A}}$ (resp. $\mathbb{P}_S^{\mathcal{A}}$) as the distribution under the null (resp. under the alternative with contaminated set S) of the joint sequence $(A^1, X^1, \ldots, A^M, X^M)$ of queries and observations. The performance of an adaptive testing procedure $(\mathcal{A}, \phi)$ is evaluated by comparing the worst-case risk

$$R(\mathcal{A}, \phi) = \mathbb{P}_{\emptyset}^{\mathcal{A}}(\phi \neq 0) + \max_{S \in \mathcal{C}} \mathbb{P}_S^{\mathcal{A}}(\phi \neq 1)$$

to the corresponding minimax risk $R_{\mathbf{AS}}^* = \inf_{\mathcal{A} \in \mathbf{AS}(M), \phi} R(\mathcal{A}, \phi)$, where the infimum is over all adaptive testing procedures $(\mathcal{A}, \phi)$ with a budget of M coordinate measurements. The minimax risk $R_{\mathbf{AS}}^*$ depends on M, although we do not write this dependence explicitly for notational ease.

Let $m = M/n$ be the equivalent number of *full vector measurements*. In the following, we will just say m *measurements* for simplicity. This change of parameters allows for easier comparison with the special case of uniform sensing, where a full vector of length n is measured m times. In particular, when $m = M/n$ is an integer, uniform sensing corresponds to the deterministic sensing procedure with $A^t = [n]$ for $t \in [m]$, $A^t = \emptyset$ for $t > m$, and $\mathbb{P}_S^{\mathcal{A}} = \mathbb{P}_S^{\otimes m}$ for $S \in \mathcal{C} \cup \{\emptyset\}$.

We are interested in the *high-dimensional* setting, where the ambient dimension n is high. All quantities such as correlation coefficient ρ, the contaminated set size k, and the number of vector measurements m will thus be allowed to depend on n. In particular, we always assume that n, k and m all go to infinity simultaneously, albeit possibly at different rates, and our main concern is to identify the range of parameters in which it is possible to construct adaptive tests whose risks converge to zero. We consider the sparse regime where $k = o(n)$. Although the case of fixed ρ is of interest, most of our results will be concerned with the case where ρ converges to zero with n. When $\rho = 1$, the problem is trivial as detecting duplicate entries in a single sample vector from the distribution allows one to perform detection perfectly, while for fixed $\rho < 1$, the problem essentially becomes easier as the measurement budget m increases.

C. Uniform Sensing and Testing

The simplest and most-natural type of non-adaptive sensing strategy we can consider is uniform sensing. As stated before, this corresponds to the choice $A^t = [n]$ for $t = 1, \ldots, m$ (recall that $m = M/n$), that is one collects m i.i.d. samples from $\mathbb{P}_S$. The minimax risk and the performance of several uniform sensing testing procedures have been analyzed in [4]. The authors of that work analyzed the performance of tests based on the *localized squared sum* statistic

$$T_{\text{loc}} = \max_{S \in \mathcal{C}} \sum_{t=1}^{m} \left(\sum_{i \in S} X_i^t \right)^2,$$

which was shown to be near-optimal in a variety of scenarios. The localized squared sum test that rejects the null hypothesis when T_{loc} exceeds a properly chosen threshold was shown to have an asymptotically vanishing risk when, for some

positive constant c,

$$\rho k \geq c \max \left(\sqrt{\frac{\log |\mathcal{C}|}{m}}, \frac{\log |\mathcal{C}|}{m} \right). \quad (2)$$

This condition was shown to be near-optimal in most regimes for the classes of k-sets and k-intervals, unless k exceeds $\sqrt{n}$. In this latter and rather easier case, the simple non-localized squared sum statistic $T_s = \sum_{t=1}^{m} \left(\sum_{i=1}^{n} X_i^t \right)^2$ is near optimal. From (2), it is easy to see that the size of the class plays an important role, as a smaller class $\mathcal{C}$ leads to a weaker sufficient condition for detection. In particular, the localized squared sum test has asymptotically vanishing risk when

$$\textbf{k-sets: } \rho \geq c \max \left(\sqrt{\frac{\log n}{km}}, \frac{\log n}{m} \right),$$

$$\textbf{k-intervals: } \rho \geq c \max \left(\frac{1}{k} \sqrt{\frac{\log n}{m}}, \frac{\log n}{km} \right).$$

Necessary conditions for detection almost matching the previous sufficient conditions have been derived in [4]. Although the dependence on the ambient dimension n is only logarithmic, this can still be significant in regimes where n is large but m is small.

D. Related Work

A closely related problem is that of detecting non zero mean components of a Gaussian vector X, referred to as the *detection-of-means* problem. This problem has received ample attention in the literature, see, for instance, [1], [7], [8], [16], [17], [22], [23] and references therein. The detection-of-means problem can be formulated as the multiple hypothesis testing problem

$$H_0 : X \sim \mathcal{N}(0, I_n),$$
$$H_1 : X \sim \mathcal{N}(\mu \mathbf{1}_S, I_n), \text{ for some } S \in \mathcal{C},$$

where $\mathbf{1}_S$ is the indicator vector of S, I_n is the identity matrix, and $\mu \neq 0$. In other words, one needs to decide whether the components of X are independent standard normal random variables or they are independent normals with unit variance, and there is a (unknown) subset S of k components that have non-zero mean. The set of contaminated components S is assumed to belong to a class $\mathcal{C}$ of subsets of $[n]$. The behavior of the minimax risk has been analyzed for various class choices $\mathcal{C}$ [1], [7], [11], [22]. Detection and estimation in this model has been analyzed under adaptive sensing in [13] and [19], where it is shown that, perhaps surprisingly, all sufficiently symmetric classes $\mathcal{C}$ lead to the same almost matching necessary and sufficient conditions for detection. This is quite different from the non-adaptive version of the problem where size and structure of $\mathcal{C}$ influence, in a significant way, possibilities of detection (see [1]).

Recall that the correlation model of Section I-A can be rewritten as

$$H_0 : U_i^t = Y_i^t, \ i \in \{1, \dots, n\},$$

$$H_1 : U_i^t = \begin{cases} Y_i^t, & i \notin S, \\ \sqrt{1-\rho} Y_i^t + \sqrt{\rho} N^t, & i \in S \end{cases}$$

for some $S \in \mathcal{C}$, with (Y_i^t), N^t independent standard normals, and that, as a consequence, the correlation model can be seen as a *random mean shift* model, with a slightly different normalization. However, most results on adaptive sensing for detection-of-means heavily hinge on the independence assumption between coordinates, which is not applicable for the detection of correlations. In particular, we shall see that the picture is more subtle in the presence of correlations.

A second problem, perhaps even more related, is that of detection in sparse principal component analysis (sparse PCA) within the *rank one spiked covariance model*, defined as the testing problem

$$H_0 : X \sim \mathcal{N}(0, I_n),$$
$$H_1 : X \sim \mathcal{N}(0, I_n + \theta u u^T),$$

for some $u \in \mathbb{R}^n$ with $\|u\|_0 = k$, $\|u\|_2 = 1$, where $\|u\|_0$ is the number of nonzero elements of u, and $\|u\|_2$ is the Euclidean norm of u. There is, also for this problem, a growing literature, see [10], [12], [26]. Note that when the coordinates of u are constrained in $\{0, 1/\sqrt{k}\}$, we recover a problem akin to that of detection of positive correlations, but with *unnormalized variances* over the contaminated set. The related problem of support estimation has been considered in [3] under the similar assumption that coordinates of u are constrained in $\left\{ 0, \pm 1/\sqrt{k} \right\}$.

E. Results and Contributions

The main contribution of this paper is to show that adaptive sensing procedures can significantly outperform the best non-adaptive tests for the model in Section I-A. We tackle the classes of k-intervals and k-sets. For k-intervals, necessary and sufficient conditions are almost matching. In particular, the number of measurements m necessary and sufficient to ensure that the risk approaches zero has almost no dependence on the signal dimension n. This is in stark contrast with the non-adaptive sensing results, where it is necessary for m to grow logarithmically with n.

For k-sets, we obtain sufficient conditions that still depend logarithmically in n, but which improve nonetheless upon uniform sensing in some regimes. Although not uniform, the proposed sensing strategy is still non-adaptive. In addition to this, in a slightly different model akin to that of sparse PCA mentioned above, we show that all previous results (both non-adaptive and adaptive) carry on, and we obtain a tighter sufficient condition for detection of k-sets, that is nearly independent of the dimension n, and also improves significantly over non-adaptive sensing. Our results are summarized in Table I. The paper is structured as follows. We obtain a general lower bound in Section II, and study various classes of contaminated sets. In Section III, we propose procedures for k-sets and k-intervals. In Section IV, we prove a tighter sufficient condition under a slightly different model, for k-sets. Finally, we conclude with a discussion in Section V.

F. Notation

We denote by $\mathbb{E}_{\mathbb{P}}$ the expectation with respect to a distribution $\mathbb{P}$. The Kullback-Leibler (KL) divergence between two

TABLE I
SUMMARY OF RESULTS (CONSTANTS OMITTED)

		reference	$\rho k \to 0$	$\rho k \to \infty$
k-sets	necessary condition	Thm. 1	$\rho k \sqrt{m} \to \infty$	-
	sufficient condition	Prop. 4	$\rho \sqrt{km} \geq \sqrt{\log \frac{n}{k}}$, and $\rho km \geq \log \frac{n}{k}$	identical
	sufficient condition (unnormalized model)	Prop. 6	$\rho \sqrt{km} \geq \log \log \frac{n}{k}$	identical
	sufficient condition (uniform, $k = o(\sqrt{n})$)	[5]	$\rho \sqrt{km} \geq \sqrt{\log n}$, and $\rho m \geq \log n$	identical
	necessary condition (uniform)	[5]	$\rho \sqrt{km} \geq \sqrt{\log \frac{n}{k^2}}$, and $\rho m \geq \log \frac{n}{k^2}$	identical
k-intervals	necessary condition	Thm. 1	$\rho k \sqrt{m} \to \infty$	-
	sufficient condition	Prop. 3	$\rho k \sqrt{m} \geq \sqrt{\log \log \frac{n}{k}}$	$\rho km \geq \log \log \frac{n}{k}$
	sufficient condition (uniform)	[5]	$\rho k \sqrt{m} \geq \sqrt{\log \frac{n}{k}}$	$\rho km \geq \log \frac{n}{k}$
	necessary condition (uniform)	[5]	$\rho k \sqrt{m} \geq \sqrt{\log \frac{n}{k}}$,	$\rho km \geq \log \frac{n}{k}$

probability distributions $\mathbb{P}$ and $\mathbb{Q}$ such that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ is $\text{KL}(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}\left[\log\left(d\mathbb{P}/d\mathbb{Q}\right)\right]$, with $d\mathbb{P}/d\mathbb{Q}$ the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$. When $\mathbb{P}$ and $\mathbb{Q}$ admit densities f and g, respectively, with respect to the same dominating measure, we write $\text{KL}(\mathbb{P} \,\|\, \mathbb{Q}) = \text{KL}(f \,\|\, g)$. We denote by $\mathbf{1}_A$ the indicator function of an event or condition A.

II. LOWER BOUNDS

We say that a sequence $z = (a^1, x^1, \ldots, a^M, x^M) \in \left(2^{[n]} \times \mathbb{R}^n\right)^M$ is *M-admissible* if $\sum_{t=1}^M |a^t| \leq M$. Consider an adaptive testing procedure $(\mathcal{A}, \phi)$, with query sequence $(A^1, \ldots, A^M) \in \left(2^{[n]}\right)^M$, and $(X^1, \ldots, X^M) \in (\mathbb{R}^n)^M$ the corresponding sequence of observations. Let $S \in \mathcal{C} \cup \{\emptyset\}$ be the set of contaminated coordinates. For $t \in [M]$, we denote by $f_{A^t \,|\, H^{t-1}}(\cdot \,|\, h^{t-1})$ the probability mass function of A^t given $H^{t-1} = h^{t-1}$, and by $f_{X^t \,|\, A^t; \, S}(\cdot \,|\, a^t)$ the density of $X^t \,|\, A^t = a^t$ over $\mathbb{R}^n$ with respect to a suitable dominating measure over $\mathbb{R}^n$ (e.g., the product of the Lebesgue measure and a point mass at 0). Therefore, the joint sequence $Z = (A^1, X^1, \ldots, A^M, X^M)$ admits a density f_S with respect to some appropriate dominating measure. For any *M*-admissible sequence $(a^1, x^1, \ldots, a^M, x^M)$, this density factorizes as

$$f_S(a^1, x^1, \ldots, a^M, x^M)$$
$$= \prod_{t=1}^M f_{A^t \,|\, H^{t-1}}(a^t \,|\, a^1, x^1, \ldots, a^{t-1}, x^{t-1}) \, f_{X^t \,|\, A^t; \, S}(x^t \,|\, a^t).$$

For concreteness, let the density f_S be zero on any joint subsequence that is not *M*-admissible. It is crucial to note that all the terms in the factorization corresponding to the sensing strategy (i.e., corresponding to the selection of A^t given the history) do not depend on S. This is central to our arguments, as likelihood ratios simplify. More precisely, for any *M*-admissible sequence $(a^1, x^1, \ldots, a^M, x^M)$,

$$\frac{f_\emptyset(a^1, x^1, \ldots, a^M, x^M)}{f_S(a^1, x^1, \ldots, a^M, x^M)} = \prod_{t=1}^M \frac{f_{X^t \,|\, A^t; \, \emptyset}(x^t \,|\, a^t)}{f_{X^t \,|\, A^t; \, S}(x^t \,|\, a^t)}$$
$$= \prod_{t=1}^M \frac{f_{X_{A^t}^t \,|\, A^t; \, \emptyset}(x_{a^t}^t \,|\, a^t)}{f_{X_{A^t}^t \,|\, A^t; \, S}(x_{a^t}^t \,|\, a^t)},$$

where the second equality follows from the sensing model.

Likelihood ratios play a crucial role in the characterization of testing performance. In particular, a classical argument (see [38, Lemma 2.6]) shows that, for any distributions $\mathbb{P}, \mathbb{Q}$ over a common measurable space Ω and any measurable function $\phi : \Omega \to \{0, 1\}$,

$$\mathbb{P}(\phi \neq 0) + \mathbb{Q}(\phi \neq 1) \geq \frac{1}{4} \exp\left(-\,\text{KL}(\mathbb{P} \,\|\, \mathbb{Q})\right).$$

Therefore

$$
\begin{aligned}
R^* &= \inf_{(\mathcal{A}, \phi)} \left[\mathbb{P}_0^{\mathcal{A}}(\phi \neq 0) + \max_{S \in \mathcal{C}} \mathbb{P}_S^{\mathcal{A}}(\phi \neq 1) \right] \\
&= \inf_{(\mathcal{A}, \phi)} \max_{S \in \mathcal{C}} \left[\mathbb{P}_0^{\mathcal{A}}(\phi \neq 0) + \mathbb{P}_S^{\mathcal{A}}(\phi \neq 1) \right] \\
&\geq \inf_{\mathcal{A}} \max_{S \in \mathcal{C}} \left[\frac{1}{4} \exp(-\,\text{KL}(\mathbb{P}_0^{\mathcal{A}} \,\|\, \mathbb{P}_S^{\mathcal{A}})) \right] \\
&= \frac{1}{4} \exp(-\sup_{\mathcal{A}} \min_{S \in \mathcal{C}} \text{KL}(\mathbb{P}_0^{\mathcal{A}} \,\|\, \mathbb{P}_S^{\mathcal{A}})).
\end{aligned}
$$

This entails that the minimax risk under adaptive sensing can be lower bounded by upper bounding the maximin KL divergence. Here, in order to bound the maximum KL divergence, we will take an approach similar to [13] for detection-of-means under adaptive sensing, although our setup differs slightly. In [13], the testing procedures measure a single coordinate at a time, while we need multiple measures per step in order to capture correlations. We have the following necessary condition.

Theorem 1: Let $\mathcal{C}$ be either the class of k-sets or k-intervals or disjoint k-intervals, and define

$$D(\rho, k) = \min\left[\frac{\rho}{2(1-\rho)}, \rho^2(k+1)\right].$$

Then the minimax risk R_{AS}^* of adaptive testing procedures with a measurement budget of $M = mn$ coordinates is lower bounded as

$$R_{AS}^* \geq \frac{\exp\left(-mkD(\rho, k)\right)}{4}.$$

As a consequence, for the risk R_{AS}^* to converge to zero, it is necessary that $mkD(\rho, k) \to \infty$.

Proof: First remark the following: for $\rho \leq 1/2$, and for any $A \subseteq [n]$,

$$\text{KL}(\mathbb{P}_0|_A \,\|\, \mathbb{P}_S|_A) \leq D(\rho, k) \,|A \cap S|.$$

The proof is given in Appendix A-B. The KL divergence between the joint probability models can we written as

$$
\begin{aligned}
& \mathrm{KL}(\mathbb{P}_0^{\mathcal{A}} \mid \mathbb{P}_S^{\mathcal{A}}) \\
&= \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[\mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[\log \frac{f_{X_{A^t}^t \mid A^t;\, \emptyset}(x_{A^t}^t \mid A^t)}{f_{X_{A^t}^t \mid A_t;\, S}(x_{A^t}^t \mid A^t)}\middle| A^t\right]\right] \\
&= \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[\mathrm{KL}(f_{X_{A^t}^t \mid A^t;\, \emptyset}(\cdot \mid A^t) \,\|\, f_{X_{A^t}^t \mid A^t;\, S}(\cdot \mid A^t))\right] \\
&= \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}[\mathrm{KL}(\mathbb{P}_0 |_{A^t} \,\|\, \mathbb{P}_S |_{A^t})] \\
&\leq D(\rho, k) \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[|A^t \cap S|\right] \\
&= D(\rho, k) \sum_{i \in S} b_i
\end{aligned}
$$

using the shorthand $b_i = \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}[\mathbf{1}_{i \in A^t}]$. Hence,

$$
\sup_{\mathcal{A}} \min_{S \in \mathcal{C}} \mathrm{KL}(\mathbb{P}_0^{\mathcal{A}} \,\|\, \mathbb{P}_S^{\mathcal{A}}) \leq D(\rho, k) \sup_{\mathcal{A}} \min_{S \in \mathcal{C}} \sum_{i \in S} b_i.
$$

Define the *class complexity*

$$
\mathfrak{C}(\mathcal{C}, M) = \sup_{\mathcal{A} \in \mathbf{AS}}\left\{\min_{S \in \mathcal{C}} \sum_{i \in S} b_i \,:\, b \in \mathbb{R}_+^n, \sum_{i=1}^{n} b_i \leq M\right\}.
$$

For any sensing strategy $\mathcal{A}$, it holds that

$$
\sum_{i=1}^{n} b_i = \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}[|A^t \cap S|] \leq M,
$$

such that

$$
\sup_{\mathcal{A}} \min_{S \in \mathcal{C}} \mathrm{KL}(\mathbb{P}_0^{\mathcal{A}} \,\|\, \mathbb{P}_S^{\mathcal{A}}) \leq D(\rho, k)\, \mathfrak{C}(\mathcal{C}, M).
$$

From [13, Lemma 3.1], we conclude that, for the both classes $\mathcal{C}_k$ and $\mathcal{D}_{[k]}$, respectively k-sets and disjoint k-intervals we have $\mathfrak{C}(\mathcal{C}_k, M) = \mathfrak{C}(\mathcal{D}_{[k]}, M) = \frac{Mk}{n} = mk$ (assuming without loss of generality for disjoint k-intervals that n/k is an integer.[1]) As $\mathfrak{C}(\cdot, M)$ is decreasing with respect to set inclusion for any fixed M, $\mathfrak{C}(\mathcal{C}_{[k]}, M) = mk$ as well, and the result follows. □

The lower bound argument in Theorem 1 yields the same lower bound for detection using any of the three classes of interest. This phenomenon is akin to what was observed in the context of detection-of-means under adaptive sensing, where the lower bounds are the same provided the classes of contaminated components are symmetric. In this setting, it was shown in addition in [13] that the condition in the lower bound is essentially sufficient and therefore, unlike in the non-adaptive counterpart of the problem, knowledge of the structure of $\mathcal{C}$ does not make the detection problem any easier. However, the problem of detection of correlations considered here seems to be more subtle in that one lacks matching upper bounds for all cases. Namely, we do not know whether:

---

[1]If n/k is not an integer, one can directly show that $\mathfrak{C}(\mathcal{D}_{[k]}, M) \leq 2mk$ and the result of the theorem for this class follows with mk replaced by $2mk$.

(a) for detection-of-correlations structure does not help; or (b) the lower bound is loose for some classes, in particular the class of k-sets.

Recall that we are interested in the characterization of the regimes for which the risk $R_{\mathbf{AS}}^*$ converges to zero as $m, k, n \to \infty$. Clearly, if ρ decays at a rate no faster than $1/k$, the previous necessary condition for the risk to vanish asymptotically is always satisfied. Nevertheless, the lower bound gives an indication about the rate at which the risk converges to zero. However, when $\rho = o(1/k)$ the situation is different, and Theorem 1 leads to the following necessary condition.

Corollary 1: Let $\mathcal{C}$ denote either the class of k-sets, k-intervals or disjoint k-intervals, and suppose $\rho = o(1/k)$. For $R_{\mathbf{AS}}^*$ to converge to zero it is necessary that $\rho k \sqrt{m} \to \infty$.

Proof: From the previous results, it is necessary that

$$
mk \min\left[\frac{\rho}{2(1-\rho)}, \rho^2(k+1)\right]
$$

goes to infinity for the risk to converge to zero. This quantity is asymptotically equivalent to $m\rho^2 k^2$, and $m\rho^2 k^2 \to \infty$ if and only if $\rho k \sqrt{m} \to \infty$. □

Recall that a sufficient condition for non-adaptive detection of k-intervals with the localized squared sum test is

$$
\rho k \sqrt{m} > c\sqrt{\log(n)} \quad \text{and} \quad \rho k m > c \log(n).
$$

When $\rho = o(1/k)$ one has, asymptotically, $\rho k < 1$ and the first condition is stronger than the second. Non-adaptive detection with k-intervals is thus possible asymptotically for $\rho k \sqrt{m} > c\sqrt{\log(n)}$. This corresponds to the condition of Corollary 1 up to a logarithmic factor in n, which implies that in the case of k-intervals, one can improve at most by a factor logarithmic in n with adaptive sensing. This can be still quite significant, and we show in Section III that this can indeed be achieved.

III. ADAPTIVE TESTS

A. The Case of k-Intervals

In this section, we study the case of the class $\mathcal{C}_{[k]}$ of intervals of length k. It is sufficient to work with the class $\mathcal{D}_{[k]}$ of disjoint intervals for the following reason: assume that one has a procedure for detection of disjoint k-intervals. Then, for detection of general k-intervals, *this procedure can be applied as if the objective was detection of disjoint $k/2$-intervals*. Indeed, if S is any k-interval, there exist at most two sets in $\mathcal{D}_{[k/2]}$ that intersect S, and at least one of them, say S', has a full intersection with S, i.e., $|S \cap S'| = k/2$. As a consequence, under mild conditions on the procedure, this leads to a sufficient condition for detection of k-intervals identical up to constants to that associated with the original procedure for disjoint k-intervals. Since up to two of the disjoint intervals can contain contaminated coordinates, the theoretical analysis still has to be slightly amended, but these technical modifications are straightforward for the methods that we propose. To keep the presentation simple, we only show how to perform detection in the case of disjoint k-intervals. Recall that $\mathcal{D}_{[k]} = \{I_1, \ldots, I_{\lfloor n/k \rfloor}\}$, where

$I_j = \{(j-1)k+1,\ldots,jk\}$ for $j \in [\lfloor n/k \rfloor]$. For simplicity, we assume that n/k is an integer. As the intervals are disjoint, the problem is equivalent to n/k independent hypothesis testing problems, each of them over vectors in $\mathbb{R}^k$ that are mutually independent. Formally, this can be cast as a testing problem over a matrix $Z \in \mathbb{R}^{\frac{n}{k} \times k}$, where Z has independent standard Gaussian entries except under the alternative where Z has a single row whose entries are mutually correlated standard Gaussian random variables with correlation ρ. In this framework, each row corresponds to one of the n/k disjoint k-intervals.

In the context of support recovery from signals with independent entries using adaptive sensing, [30], [31] have proposed the sequential thresholding (ST) procedure, which is based on an intuitive bisection idea. Although initially introduced for support estimation, ST can be easily adapted to detection, and we present such results here. In addition, we present a slight generalization to signals with independent *vector entries*, which will allow us to apply the modified procedure to the disjoint k-intervals problem. We will also use the original ST procedure in Section IV-B, and for this reason, we first present the method using general notations here. Let $\mathbb{Q}_0$ and $\mathbb{Q}_1$ be two probability distributions over $\mathbb{R}^{\tilde{d}}$, and let $Z \in \mathbb{R}^{\tilde{n} \times \tilde{d}}$ be a random matrix. Consider the multiple testing problem defined as follows. Under the null, Z has rows identically distributed according to $\mathbb{Q}_0$. Under the alternative, a small unknown subset of $\tilde{k}$ rows of Z are distributed according to $\mathbb{Q}_1$, while the remaining rows are distributed according to $\mathbb{Q}_0$. In both cases, all rows are independent. More formally, denote by $Z_1, \ldots, Z_{\tilde{n}}$ the rows of Z, such that the testing problem is

$$H_0 : Z \sim \mathbb{Q}_0^{\otimes \tilde{n}},$$
$$H_1 : Z_i \sim \mathbb{Q}_0 \quad \text{for } i \notin S, \quad Z_i \sim \mathbb{Q}_1 \text{ for } i \in S,$$

for some $S \in \mathcal{C}$ with $|S| = \tilde{k}$, where, as already mentioned, all rows are independent in both cases. We refer to this testing problem as that of detection from *signals with independent (vector) entries*. The framework of adaptive sensing introduced in Section I-B can be easily adapted to this model. In this case, in order to allow for vector entries, we consider that the experimenter is allowed to obtain samples from rows of Z, and that he can select which rows to query in a sequential manner as previously, under the constraint that the total number of rows measured be less than M. We also refer to this straightforward extension as adaptive sensing, and we say that $\tilde{m} = M/\tilde{n}$ is the number of *measurements* (i.e., $\tilde{m}$ is the equivalent number of times the full matrix Z was observed).

Sequential thresholding is a procedure for testing with adaptive sensing within the type of model just mentioned. Assume that $\mathbb{Q}_0$ and $\mathbb{Q}_1$ admit densities f_0 and f_1, respectively, with respect to some common dominating measure, and for $i \in [n]$, denote by

$$LR(f_1|f_0; z_i^1, \ldots, z_i^{\tilde{m}}) = \frac{\prod_{t=1}^{\tilde{m}} f_0(z_i^t)}{\prod_{t=1}^{\tilde{m}} f_1(z_i^t)}$$

the likelihood ratio associated to i.i.d. observations $z_i^1, \ldots, z_i^{\tilde{m}} \in \mathbb{R}^{\tilde{d}}$ of Z_i, the i-th row of Z. ST proceeds
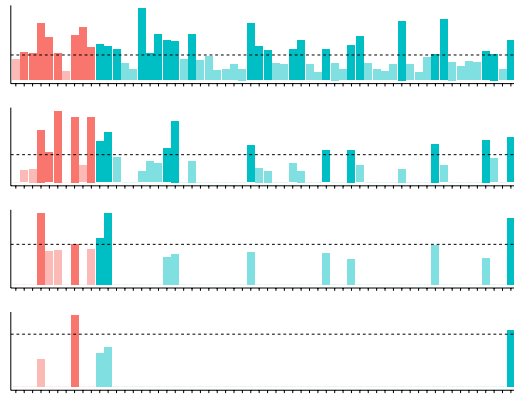

Fig. 2. Illustration of sequential thresholding with $k = 10$, $n = 60$: contaminated coordinates are the first ten on the left. Bars depict likelihood ratios associated with each coordinate: at each step, coordinates with likelihood ratio below a threshold are thrown away. First step shown in top row, last step shown in bottom row.

Input: $K = \lfloor \log_2(\tilde{n}) \rfloor$ (number of steps),
$\quad \bar{m} = \frac{\tilde{m}}{4}$,
$\quad \gamma = \text{median}_{z_1^1,\ldots,z_1^{\bar{m}} \sim f_0}(LR(f_1|f_0; z_1^1, \ldots, z_1^{\bar{m}}))$
(threshold) [2]
Initialization: $\mathcal{S}_0 = \{1, \ldots, \tilde{n}\}$
for all $r = 1, \ldots, K$ **do**
$\quad$ **for all** $i \in \mathcal{S}_{r-1}$ **do**
$\quad\quad$ measure $z_i^1, \ldots, z_i^{\bar{m}} \sim Z_i$
$\quad\quad$ compute $LR_i = LR(f_1|f_0; z_i^1, \ldots, z_i^{\bar{m}})$
$\quad$ **end for**
$\quad$ $\mathcal{S}_r = \{i \in \mathcal{S}_{r-1} : LR_i > \gamma\}$
$\quad$ **if** $\sum_{r=0}^{K} |\mathcal{S}_r| > \tilde{n}$ **then**
$\quad\quad$ **return** no detection
$\quad$ **end if**
end for
return detection if $\mathcal{S}_K \neq \emptyset$

Fig. 3. Sequential thresholding procedure.

as outlined in Figure 3. Initially, ST measures all $\tilde{n}$ rows $\bar{m} = \tilde{m}/4$ times, and throws away a fraction (of about half under the null) of the $\tilde{n}$ rows based on the values of the likelihood ratios. This is repeated with the remaining rows a number of times logarithmic in $\tilde{n}$, at which point ST calls detection if some coordinates have not been thrown away. This is illustrated in Figure 2.

The following result is easily deduced from the analysis of ST for support estimation.

Proposition 1: [Sufficient Condition for ST] Assume $\tilde{k}/\tilde{n} \to 0$, and

$$\liminf_{\tilde{n} \to \infty} \frac{\tilde{m} \, \text{KL}(f_0 \,||\, f_1)}{4 \log \log_2 \tilde{n}} > 1,$$

then the sequential thresholding procedure with a budget of $\tilde{m}$ measurements has risk tending to zero as $\tilde{n}$ goes to infinity.

---

[1]Here, $z_1^1, \ldots, z_1^{\tilde{m}}$ denote without loss of generality observations of the first row, as rows are exchangeable under the null.

Proof: We begin by showing that the event of termination upon $\sum_{r=0}^{K} |\mathcal{S}_r| > \tilde{n}$ has an asymptotically vanishing probability. Assume the alternative hypothesis with contaminated set S. Then, similarly as in [13, Proposition 4.1], using Bernstein's inequality for sums of truncated hypergeometric variables,

$$P\left(\sum_{r=0}^{K} |\mathcal{S}_r| > \tilde{n}\right) \leq \exp\left(-\frac{\tilde{n}/4 - \tilde{k}}{4 + \frac{2K}{3}}\right),$$

which converges to zero. The application of the Chernoff-Stein lemma as in [30] allows us to bound the probability of error as follows. The type I error of the procedure is bounded by

$$\frac{\tilde{n} - \tilde{k}}{2^K}.$$

Let $E_{i,t}$ denote the event that the likelihood ratio is below γ for coordinate i at step t (in which case, coordinate i will not be included in $\mathcal{S}_t$). Without loss of generality, assume that $1 \in S$. The type II error is

$$\mathbb{Q}_1\left(\cap_{i \in S} \left(\cup_{t=1}^{K} E_{i,t}\right)\right) \leq \left(K\mathbb{Q}_1\left(E_{1,1}\right)\right)^{\tilde{k}}.$$

We write $a \doteq e^{-\bar{m}D}$ for $\lim_{\bar{m} \to \infty} \frac{\log a}{\bar{m}} = D$. From the Chernoff-Stein lemma,

$$\mathbb{Q}_1\left(E_{1,1}\right) \doteq e^{-\bar{m}\,\mathrm{KL}(f_0 \| f_1)}.$$

Hence, for $K = (1 + \varepsilon_1) \log_2 n$ and $\varepsilon_2 > 0$, there exists $\bar{m}_0$ such that for $\bar{m} \geq \bar{m}_0$, the type II error is bounded by

$$\left(Ke^{-\bar{m}(\mathrm{KL}(f_0 \| f_1) - \varepsilon_2)}\right)^{\tilde{k}}$$

$$= \exp\left(\tilde{k} \log\left[(1 + \varepsilon_1) \log_2 n\right] - \bar{m}\tilde{k}(\mathrm{KL}(f_0 \| f_1) - \varepsilon_2)\right).$$

Hence, the risk goes to zero if for some $\varepsilon_1, \varepsilon_2 > 0$, it holds that

$$\liminf_{\tilde{n} \to \infty} \frac{\bar{m}(\mathrm{KL}(f_0 \| f_1) - \varepsilon_2)}{\log\left[(1 + \varepsilon_1) \log_2 n\right]} > 1.$$

As a consequence, for the risk to go to zero, it is sufficient that

$$\liminf_{\tilde{n} \to \infty} \frac{\bar{m}\,\mathrm{KL}(f_0 \| f_1)}{\log \log_2 n} > 1.$$

The result follows by substituting $\bar{m}$ with $\frac{\tilde{m}}{4}$. $\square$

Note that the ST procedure does not require knowledge of $\tilde{k}$. ST can be applied to the case of k-intervals, as we demonstrate in the next section.

We now show how the previous procedure can be used for adaptive detection with disjoint k-intervals. As before, we assume that n/k is an integer. Define $\tilde{n} = n/k$, $\tilde{k} = 1$, $\tilde{m} = m$, and $\tilde{d} = k$. Let $\mathbb{Q}_0 = \mathbb{P}_0|_{I_1}$ be the joint probability distribution over an interval under the null, and $\mathbb{Q}_1 = \mathbb{P}_S|_S$ be the joint probability distribution over the contaminated interval under the alternative with contaminated interval $S \in \mathcal{D}_{[k]}$. Here, the choice of the interval used in $\mathbb{Q}_0$ does not matter, as intervals are exchangeable under the null hypothesis. We refer to the corresponding sequential thresholding procedure as *ST for disjoint k-intervals*. This procedure is illustrated in Figure 4. This provides the following sufficient condition for detection of disjoint k-intervals.
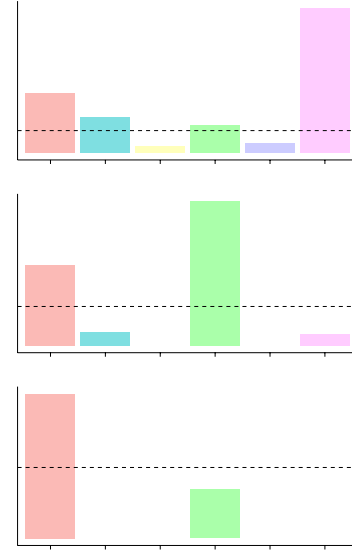

Fig. 4. Illustration of sequential thresholding for k-intervals, with $n/k = 6$ intervals of size k. Bars depict likelihood ratios associated with the intervals.

Proposition 2: Assume that ρ converges to zero. There exists numerical constants C_3 and C_4 such that, when either

$$\rho k \to \infty \quad and \quad m \log(1 + \rho k) \geq C_3 \log \log(n/k),$$

or

$$\rho k \to 0 \quad and \quad \rho k \sqrt{m} \geq C_4 \sqrt{\log \log(n/k)},$$

the sequential thresholding procedure for disjoint k-intervals has risk converging to zero.

Proof: The detailed computations can be found in Appendix A-C. Assume that $\rho k > 1$, then

$$\mathrm{KL}(\mathbb{Q}_0 \| \mathbb{Q}_1) \geq \frac{\log(1 + \rho k)}{10}.$$

Similarly, when $\rho k < 1/2$ and $k > 32$,

$$\mathrm{KL}(\mathbb{Q}_0 \| \mathbb{Q}_1) \geq \frac{\rho^2 k^2}{16}.$$

Combined with Proposition 1, this gives the desired result. $\square$

Consider the case where $\rho k \to \infty$. In that case, omitting constant factors, sequential thresholding would succeed for $m \geq \frac{\log \log(n)}{\log(1 + \rho k)}$. Recall that uniform non-adaptive testing is possible for $m \geq \frac{c \log n}{\rho k}$. When $\rho k > \log(n)$ asymptotically, both conditions are trivially satisfied for m constant, while when $\rho k < \log(n)$, we already improve upon non-adaptive tests. In spite of this, the dependence on ρk of our sufficient condition when $\rho k \to \infty$ is logarithmic, while it is only linear for $\rho k \to 0$. This may appear surprising, as one may argue the former case corresponds to a regime where the signal is stronger (and so the problem should be easier). However, this surprising fact is solely an artifact from the sequential thresholding procedure, and from the fact that ST does not require knowledge of k. This results in a sufficient condition that is independent of k. In particular, it does not become easier to satisfy as k increases, but it can be fixed through a small
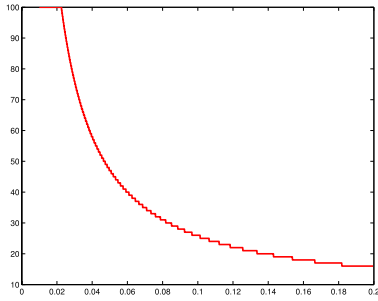
Fig. 5.   Optimal p as a function of ρ, for $k = 100$.

modification of the sensing methodology that we present in the following.

In order to recover the same linear dependence in both cases, we propose to add a subsampling stage prior to sequential thresholding. This subsampling can be decided before any data is collected, and thus can be viewed as a non-adaptive aspect of the entire procedure. Consider the simple deterministic subsampling scheme wherein one keeps the first p coordinates per interval, for some $p \in \{2, \dots, k\}$, and measures each p-tuple $\left\lfloor \frac{mn}{pn/k} \right\rfloor = \left\lfloor \frac{mk}{p} \right\rfloor$ times. This prompts the following question: is there a value of p that allows one to detect more easily? Define the p-truncated intervals as $I_j^p = \{(j - 1)k + 1, \dots, (j - 1)k + p\}$ for $j \in [n/k]$. Formally, we consider the deterministic sensing strategy $\mathcal{A}_p = (A^t)$ where for $t \in \left[\left\lfloor \frac{mk}{p} \right\rfloor \right]$,

$$A^t = \bigcup_{j \in [n/k]} I_j^p.$$

As this involves one simple testing problem per interval, the difficulty of testing is essentially characterized by the KL divergence $\mathrm{KL}(\mathbb{P}_0^{\mathcal{A}_p} \| \mathbb{P}_S^{\mathcal{A}_p})$ between the distributions under the null and the alternative. In this section, we make explicit the dependence of $\mathbb{P}_S$ on p by using the notation $\mathbb{P}_S^p$. Consider any fixed $S \in \mathcal{D}_{[k]}$, then the best KL divergence that can be obtained is

$$\max_{p \in \{2, \dots, k\}} \mathrm{KL}\left(\mathbb{P}_0^{\mathcal{A}_p} \| \mathbb{P}_S^{\mathcal{A}_p} \right) = \max_{p \in \{2, \dots, k\}} \sum_{t=1}^{\left\lfloor \frac{mk}{p} \right\rfloor} \mathrm{KL}(\mathbb{P}_0^p \| \mathbb{P}_S^p)$$
$$= \left\lfloor \frac{mk}{p} \right\rfloor \max_{p \in \{2, \dots, k\}} \mathrm{KL}(\mathbb{P}_0^p \| \mathbb{P}_S^p),$$

which is independent of S. Due to nonlinearity in the KL divergence the optimal value of p is generally different than k, as illustrated in Figure 5. The optimal p and corresponding optimal value seem hard to compute analytically, but numerical evidence shows that, for ρ away from zero, the optimal p is of the order of ρ^{-1}. This observation is sufficient for our purposes, and is formalized below. Remark that when $\rho k < 1$, the optimal value of p is clamped to k.

Equipped with this subsampling stage when $\rho k \to \infty$, we can now modify the ST for k-intervals procedure as follows: when $\rho k \to \infty$, set $\tilde{m} = \left\lfloor \frac{mk}{p} \right\rfloor$, $\tilde{d} = \left\lceil \frac{1}{\rho} \right\rceil$, and use

only observations corresponding to $\tilde{d}$ coordinates per interval. We refer to this new procedure as the *modified sequential thresholding for disjoint k-intervals*.

Proposition 3: Assume that ρ converges to zero. There exists numerical constants C_5 and C_6 such that, when either

$$\rho k \to \infty \quad and \quad \rho km \geq C_5 \log \log(n/k),$$

or

$$\rho k \to 0 \quad and \quad \rho k \sqrt{m} \geq C_6 \sqrt{\log \log(n/k)},$$

the modified sequential thresholding procedure for disjoint k-intervals has risk converging to zero.

Proof: We have the following straightforward new lower bound: with $p = \left\lceil \frac{1}{\rho} \right\rceil$, when $\rho k > 1$, we have $\left\lceil \frac{1}{\rho} \right\rceil < k + 1$, and as a consequence,

$$\mathrm{KL}(\mathbb{P}_0^p \| \mathbb{P}_S^p) \geq \frac{\log 2 - 1/2}{2} \geq \frac{1}{11}.$$

Although the lower bound appears weaker than previously, this corresponds to a setting where more measurements can be carried out. The sufficient condition for ST leads to the result.   □

The adaptive procedure allows us to obtain a mild dependence on the original dimension n of the problem. When $\rho = o(1/k)$, this sufficient condition almost matches the lower bound of Corollary 1, while when $\rho k \to \infty$, the sufficient condition is already satisfied for $m = \log \log(n/k)$.

B. The Case of k-Sets: Randomized Subsampling

In this section, we consider the class $\mathcal{C}_k$ of k-sets. In this case, we do not currently know whether a procedure along the lines of ST can be successfully applied. However, the idea of subsampling the coordinates can still be used to yield modest but important performance gains. While for disjoint k-intervals a deterministic subsampling was sufficient, this is not the case for k-sets, where any deterministic subsampling that selects less than about $n-k$ coordinates cannot have risk converging to zero. For this reason, we consider a *randomized* subsampling of the coordinates.

Consider a sample B of $\left\lfloor \frac{2np}{k} \right\rfloor$ elements drawn without replacement from $[n]$ for some $p \geq 2$. Let $\theta : \mathbb{R}^{\lfloor 2np/k \rfloor} \to \{0, 1\}$ be the localized squared sum test with ambient dimension $\left\lfloor \frac{2np}{k} \right\rfloor$, and contaminated sets $\mathcal{C} = \mathcal{C}_{\lfloor p \rfloor}$ of size $\lfloor p \rfloor$, and consider the sensing strategy defined by

$$A^1 = \dots = A^{\left\lfloor \frac{mk}{2p} \right\rfloor} = B.$$

We refer to the adaptive sensing procedure $((A^t), \theta)$ as the *randomized testing procedure*. Define $Y = |B \cap S|$ (resp. $Y = 0$) under the alternative with contaminated $S \in \mathcal{C}_k$ (resp. under the null), which is the number of contaminated elements in the subsample. Clearly Y is a hypergeometric random variable with expectation $\frac{k}{n} \left\lfloor \frac{2n}{k} p \right\rfloor \in [2p - k/n, 2p]$. In words, we consider a subsample of the coordinates, with about $2p$ contaminated coordinates (in expectation) under the

alternative, and we apply the (non-adaptive) localized squared sum test.

Note that the procedure is strictly non-adaptive, as the subsampling can be decided in advance. However, this sensing strategy is a bit different than uniform sensing, as not all coordinates are measured. Nonetheless, this allows one to detect under weaker conditions than with uniform non-adaptive sensing when k is large enough.

Proposition 4: Let $2 \le p \le k$ such that p goes to infinity. Assume that ρ converges to zero and that

$$\rho m k \ge \frac{C_1 \log \frac{2pn}{k}}{\left[1 - \frac{1}{m} - \frac{1}{k}\right]}, \quad \text{and} \quad \rho \sqrt{mk} \ge \frac{C_1 \sqrt{\log \frac{2pn}{k}}}{\sqrt{1 - \frac{1}{m} - \frac{1}{k}}},$$

for some constant C_1, then the randomized testing procedure has risk converging to zero.

Proof: Let η_I (resp. η_{II}) be the risk of type I (resp. of type II) for θ. The type I error of the randomized testing procedure is $p_I = \eta_I$. Let $p_+ = P(Y \ge \lfloor p \rfloor)$ the probability of the sample containing at least $\lfloor p \rfloor$ contaminated elements, and $p_- = 1 - p_+$. Note that since $\frac{2np}{k}\frac{k}{n} = 2p$ goes to infinity, we can assume that Y is distributed according to a Poisson distribution with parameter $2p$, as this is asymptotically equivalent to the hypergeometric distribution. Hence, we have $p_- = P(Y < \lfloor p \rfloor) \le \left(1 + \frac{p(2p)^p}{p!}\right)\exp(-2p)$. Using $p! \ge \sqrt{2\pi p}\left(\frac{p}{e}\right)^p$, we have that $p_- \le \exp(-2p) + \sqrt{p}\exp(-p/4)$, which converges to zero. The type II error of the randomized testing procedure is $p_{II} = p_+\eta_{II} + p_-(1 - \eta_I) \le \eta_{II} + p_-$. It remains to show that η_I and η_{II} both go to zero. This follows from the sufficient conditions for the localized squared sum test, and from $\lfloor p \rfloor \left\lfloor \frac{mk}{2p} \right\rfloor \ge \frac{mk}{2}\left[1 - 1/p + \frac{2(1-p)}{mk}\right] \ge \frac{mk}{2}\left[1 - 1/p - 1/m\right]$. Hence, the sufficient conditions for the localized squared sum test θ provides the result. $\square$

In particular, for $p = \log\log n$, it is sufficient that, omitting constants,

$$\rho m k \ge \log\frac{n}{k}, \quad \rho\sqrt{mk} \ge \sqrt{\log\frac{n}{k}},$$

to ensure the detection risk converges to zero. This does not match the adaptive lower bound, and the dependence on n is still logarithmic. However, this already improves upon the setting of uniform non-adaptive sensing when $k \ge \frac{m}{\log n}$. Indeed, recall that using uniform sensing, the sufficient condition is

$$\rho m \ge \log n, \quad \rho\sqrt{mk} \ge \sqrt{\log n}.$$

The first condition is insensitive to subsampling, due to the dependence in mk, and we do not improve with respect to it. The second condition, however, only depends on m, and does not get easier to satisfy when k is large. Hence, our result shows that it is more efficient when k is large enough to reduce to a problem with an almost constant contaminated set size, but with an increased budget of full vector measurements.

IV. UNNORMALIZED CORRELATION MODEL

A. Model and Extensions of Previous Results

An alternative choice to the previous correlation model is the following *unnormalized model* with covariance matrix

$$(\bar{\Sigma}_S)_{i,j} = \begin{cases} 1, & i = j, \ i \notin S, \\ 1 + \rho, & i = j, \ i \in S, \\ \rho, & i \ne j, \text{ and } \ i,j \in S, \\ 0 & \text{otherwise.} \end{cases}$$

under the alternative with contaminated set $S \in \mathcal{C}$. This model is a special case of the *rank one spiked covariance* model introduced in [25]. Observe that this correlation model can also be rewritten as

$$H_0 : X_i^t = Y_i^t, \ i \in \{1, \dots, n\},$$
$$H_1 : X_i^t = \begin{cases} Y_i^t, & i \notin S, \\ Y_i^t + \sqrt{\rho}N^t, & i \in S \end{cases} \quad \text{for some } S \in \mathcal{C},$$

with (Y_i^t), N^t independent standard normals. This can thus be interpreted as a random additive noise model, as for the model of Section I-A. Observe that our original correlation detection model is obtained by normalizing each component such that the components have unit variance. This is a minor difference that does not essentially change the difficulty of detection in the non-adaptive setting (indeed all upper and lower bounds proved in [4] can be reproved for this model with minor modifications). Interestingly, however, under adaptive sensing the information provided by the higher variance in the contaminated components can be exploited to give a major improvement over the normalized model. This may be done by applying the sequential thresholding algorithm to the squares of the components as described below.

In the following, for any quantity X relative to the *normalized* model of Section I-A, we denote by $\bar{X}$ the corresponding quantity related to the unnormalized model. All of previous results can be shown to hold for this model as well. As already mentioned, this includes the necessary and sufficient conditions of [4] (Proposition 10 in Appendix), but also the lower bound of Theorem 1 (Proposition 11 in Appendix), and sufficient conditions for k-sets and k-intervals of Propositions 4 and 3 (Proposition 13 in Appendix). In particular, the procedures associated to the sufficient conditions can be used with little modifications.

B. The Case of k-Sets

The procedure proposed below combines randomized subsampling with sequential thresholding, in order to capitalize on the unnormalized model. Consider the second moments $Y_i = X_i^2$. Under the alternative with contaminated set $S \in \mathcal{C}$, Y_i is distributed as follows: (a) for $i \notin S$, Y_i is distributed according to a chi-squared distribution with one degree of freedom (that we denote by χ_1^2), (b) for $i \in S$, Y_i is distributed as $(1+\rho)\chi_1^2$. Note that under our sensing model, it is perfectly legitimate to sample $A_1 = \{1\}, \dots, A_n = \{n\}$, and thus obtain independent samples of each of the coordinates of the random vector. In particular, this allows us to obtain independent

samples from the coordinates of Y. As a consequence, we can directly apply ST to detect increased variance over a subset of the coordinates.

As already mentioned, ST does not require knowledge of k, which results in a sufficient condition that is independent of k. This condition can, however, be significantly weakened using the random subsampling used in last section. As in Proposition 4, this is due to the fact that by subsampling, one can increase the budget of full vector measurements, while the decrease in the contaminated set size does not impact the sufficient condition for detection. This is summarized in the following result, which can be proved similarly as Proposition 4.

Proposition 5 (Sufficient Condition for ST+Randomized Subsampling): Assume $\tilde{k}/\tilde{n} \to 0$, and

$$\liminf_{\tilde{n}\to\infty} \frac{\tilde{m}\tilde{k}\, \mathrm{KL}(f_0 \,||\, f_1)}{(\log\log_2 \tilde{n})^2} > 1,$$

then the sequential thresholding procedure with randomized subsampling ($p = \log\log_2 \tilde{n}$) and a budget of $4\tilde{m}$ full vector measurements has risk tending to zero as $\tilde{n}$ goes to infinity.

Let $\tilde{n} = n$, $\tilde{k} = k$, and $\tilde{m} = m$. Let $\mathbb{Q}_0$ be the χ_1^2 distribution, and $\mathbb{Q}_1$ be the $(1 + \rho)\chi_1^2$ distribution, both with respect to Lebesgue's measure. We consider the associated sequential thresholding procedure (with randomized subsampling), with the previous modification of sampling independent single coordinates. We refer to this procedure as *variance thresholding*. This leads to the following sufficient condition for detection.

Proposition 6: Assume that ρ converges to zero and that

$$\rho\sqrt{km} \geq C_2 \log\log_2 n$$

for some constant C_2. Then, the risk of the variance thresholding procedure converges to zero.

Proof: Let g be the density of a χ_1^2-distributed random variable, such that the density of a $(1 + \rho)\chi_1^2$-distributed random variable is given by $\frac{1}{1+\rho}g\left(\frac{\cdot}{1+\rho}\right)$. Then, using $g(x) \propto x^{-1/2}e^{-x/2}$,

$$\mathrm{KL}(\chi_1^2 \,||\, (1 + \rho)\chi_1^2)$$

$$= \int_{\mathbb{R}} \log\left(\frac{g(x)}{\frac{1}{1+\rho}g\left(\frac{x}{1+\rho}\right)}\right) g(x)dx$$

$$= \log(1 + \rho) + \int_{\mathbb{R}} \log\left(\frac{x^{-1/2}e^{-x/2}}{\left(\frac{x}{1+\rho}\right)^{-1/2}e^{\frac{-x}{2(1+\rho)}}}\right) g(x)dx$$

$$= \log(1 + \rho) + \int_{\mathbb{R}} \log\left(\frac{e^{\frac{-\rho x}{2(1+\rho)}}}{(1+\rho)^{1/2}}\right) g(x)dx$$

$$= \frac{\log(1 + \rho)}{2} - \frac{\rho}{2(1+\rho)} \int_{\mathbb{R}} xg(x)dx.$$

As the expectation of a χ_1^2-distributed random variable is one, this leads to

$$\mathrm{KL}(\chi_1^2 \,||\, (1 + \rho)\chi_1^2) = \frac{1}{2}\left[\log(1 + \rho) - \frac{\rho}{1 + \rho}\right]$$

$$= \frac{\rho^2}{4} + o(\rho^2).$$

Plugging this expression into the sufficient condition of Proposition 5 provides the result. $\qquad\square$

Assume for the following discussion that $\rho k \to 0$. The necessary condition that we have established previously is that $\rho k\sqrt{m}$ goes to infinity. Neglecting the double log factor, the sufficient condition that we have just obtained is that $\rho\sqrt{km}$ goes to infinity, which is stronger. Hence, there is a gap between the sufficient and necessary condition. In particular, that $\rho k\sqrt{m}$ goes to infinity was shown to be near-sufficient for detection with k-intervals, and the gap that we observe for k-sets does not allow us to conclude as to whether structure helps for detection (as is the case under non-adaptive sensing).

Recall that the unnormalized model is similar to that of detection in the problem of sparse PCA. The method of *diagonal thresholding* (also referred to as *Johnstone's diagonal method*) is a simple and tractable method for detection (and support estimation) in sparse PCA (with uniform non-adaptive sensing), which consists in testing based on the diagonal entries of empirical covariance matrix - that is, the empirical variances. Hence, it is similar to the method that we consider here, except that we estimate variances based on independent samples for each coordinate. Note that this last point is essential to our method. Indeed, consider the opposite case where we do not use independent samples for each coordinates. For the sake of illustration, assume $\rho = 1$, such that the contaminated components are exactly equal. In this case, the probability of throwing away one component is equal to that of throwing away *all* contaminated components, and failure will occur with fixed non small probability due to the use of dependent samples.

Finally, it is noteworthy that a naïve implementation of the optimal test in the non-adaptive setting has complexity $O(n^k)$, while with adaptive sensing, we obtain a procedure that can be carried out in time and space linear in n, and still improves significantly with respect to the non-adaptive setting.

V. DISCUSSION

We showed that for k-intervals, adaptive sensing allows one to reduce the logarithmic dependence in n of sufficient conditions for non-adaptive detection to a mild $\log\log n$, and that this is near-optimal in a minimax sense.

For k-sets, the story is less complete. The sufficient condition obtained in the unnormalized model is still stronger than the sufficient condition obtained for k-intervals, and does not match our common lower bounds, which leaves open the question of *whether structure helps under adaptive sensing for detection of correlations*? The analogous question for detection-of-means has a negative answer, meaning structure does not provide additional information for detection. However, for detection-of-correlations a definite answer is still elusive. Another open question is to what extent adaptive sensing allows one to overcome the exponential computational complexity barrier that one can encounter in the non-adaptive setting.

Aside from the normalized and unnormalized correlation models, other types of models can be considered. A more general version of our normalized model has been analyzed

in [4], where the correlations need not be all the same, leading to results that involve the mean correlation coefficient $\rho_{\text{avg}} = \left(\sum_{i,j \in S : i \neq j} (\Sigma_S)_{i,j}\right)/k(k-1)$. In addition, we assume in most procedures that ρ and/or k are known, and it would be of interest to have procedures that do not require such knowledge.

APPENDIX A
PROOFS AND COMPUTATIONS

A. Inequalities and KL Divergences

In this section, we collect elementary inequalities that we use repeatedly in the computations.

$$\text{For } x > -1, \quad \log(1+x) \leq x, \tag{3}$$

$$\text{For } x > 0, \quad \log(1+x) + \frac{1}{1+x} - 1 \leq x^2, \tag{4}$$

$$\text{For } 0 < x < 1/2, \quad \log(1-x) + \frac{1}{1-x} - 1 \leq 2x^2, \tag{5}$$

$$\text{For } x < 1, \quad -\log(1-x) - \frac{1}{1-x} + 1 \leq x^2, \tag{6}$$

$$\text{For } x \in]-1,1], \quad \log(1+x) + \frac{1}{1+x} - 1 \geq \frac{x^2}{8}, \tag{7}$$

$$\text{For } x \geq 1, \quad \log(1+x) \tag{8}$$

$$+ \frac{1}{1+x} - 1 \geq \frac{\log(1+x)}{5}. \tag{9}$$

The following expression of the KL divergence is used throughout the paper.

Proposition 7: We have

$$\text{KL}(\mathbb{P}_0 \,\|\, \mathbb{P}_S) \tag{10}$$

$$= \frac{\mathbf{I}_{k \geq 2}}{2}\left[k\left(-1 + \frac{1}{1-\rho} + \log(1-\rho)\right)\right. \tag{11}$$

$$- \left(\frac{1}{1-\rho} + \log(1-\rho)\right)$$

$$\left. + \left(\frac{1}{1+\rho(k-1)} + \log(1+\rho(k-1))\right)\right]. \tag{12}$$

Proof: The KL divergence between $\mathbb{P}_0$ and $\mathbb{P}_S$ can be computed using the standard formula for KL divergence between two centered Gaussian vectors, with covariance matrices

$$\Sigma_0 = I_n, \quad \Sigma_1 = \Sigma_S.$$

When $k < 2$, the divergence is zero, and we will thus assume $k \geq 2$. Up to a simultaneous permutation of rows and columns,

$$\Sigma_S = \begin{bmatrix} I_{n-k} & \\ & J_\rho(k) \end{bmatrix}$$

where $J_\rho(k) \in \mathbb{R}^{k \times k}$ has unit diagonal and coefficients equal to ρ everywhere else. $J_\rho(k)$ is a symmetric matrix, hence diagonalizable, and has eigenvalues $1 - \rho$ with multiplicity $k-1$ and $1+(k-1)\rho$ with multiplicity one. As a consequence, we have, for $k \geq 2$,

$$\log \det \Sigma_S = (k-1)\log(1-\rho) + \log(1+\rho(k-1))$$

$$\text{trace } \Sigma_S^{-1} = (n-k) + \frac{k-1}{1-\rho} + \frac{1}{1+\rho(k-1)}.$$

The KL divergence is thus

$$\text{KL}(\mathbb{P}_0 \,\|\, \mathbb{P}_S)$$

$$= \frac{1}{2}\left[\text{trace}(\Sigma_1^{-1}\Sigma_0) - n - \log(\det \Sigma_0 / \det \Sigma_1)\right]$$

$$= \frac{1}{2}\left[(n-k) + \frac{k-1}{1-\rho} + \frac{1}{1+\rho(k-1)} - n\right.$$

$$\left. + (k-1)\log(1-\rho) + \log(1+\rho(k-1))\right]$$

$$= \frac{1}{2}\left[k\left(-1 + \frac{1}{1-\rho} + \log(1-\rho)\right)\right.$$

$$- \left(\frac{1}{1-\rho} + \log(1-\rho)\right)$$

$$\left. + \left(\frac{1}{1+\rho(k-1)} + \log(1+\rho(k-1))\right)\right]. \qquad \square$$

B. Proof of Bound on KL Divergence

Proof: First note since the KL divergences are independent of n, it is sufficient to use the expressions of Proposition 7 with a contaminated set of size $s = |A \cap S| \leq k$. As previously, we assume $s \geq 2$, as the result is trivial otherwise. Consider the expression for the KL divergence given in (10). Using (3), we obtain

$$\text{KL}(\mathbb{P}_0|_A \,\|\, \mathbb{P}_S|_A)$$

$$= \text{KL}(\mathbb{P}_0 \,\|\, \mathbb{P}_{S \cap A})$$

$$\leq \frac{1}{2}\left[s\left(-1 + \frac{1}{1-\rho} + \log(1-\rho) + \rho\right)\right.$$

$$- \left(\frac{1}{1-\rho} + \log(1-\rho)\right) + \left(\frac{1}{1+\rho} - \rho\right)\right]$$

$$= \frac{1}{2}\left[s\left(\rho + \frac{\rho}{1-\rho} + \log(1-\rho)\right)\right.$$

$$\left. + \frac{-2\rho}{1-\rho^2} - \log(1-\rho) - \rho\right] \leq \frac{\rho s}{2(1-\rho)}.$$

Using (4) and (6), we obtain

$$\text{KL}(\mathbb{P}_0 \,\|\, \mathbb{P}_S) \leq \frac{1}{2}\left[(s-1)^2\rho^2 + 2s\rho^2 + \rho^2\right]$$

$$= \frac{\rho^2}{2}\left[(s-1)^2 + 2s + 1\right]$$

$$\leq \frac{\rho^2 s(k+1)}{2}. \qquad \square$$

C. Proof of Proposition 2

Proof: We have $\text{KL}(\mathbb{Q}_0 \,\|\, \mathbb{Q}_1) = kf(\rho) + h(\rho)$ with

$$f(\rho) = \frac{1}{2}\left[(1-\rho)^{-1} + \log(1-\rho) - 1\right],$$

$$h(\rho) = \frac{1}{2}\left[-\left(\frac{1}{1-\rho} + \log(1-\rho)\right)\right.$$

$$\left. + \left(\frac{1}{1+(p-1)\rho} + \log(1+(p-1)\rho)\right)\right].$$

As previously, using (7), $f(\rho) \geq \frac{\rho^2}{16}$. Assume that $\rho k < 1$ and $k > 7$, then using (5) and (7),

$$
\begin{aligned}
\mathrm{KL}&(\mathbb{Q}_0 \,||\, \mathbb{Q}_1) \\
&\geq \frac{\rho^2 k}{16} + h(\rho) \\
&\geq \frac{\rho^2 k}{16} - \frac{1}{2}\left[1 + 2\rho^2\right] + \frac{1}{2}\left[1 + \frac{\rho^2(k-1)^2}{8}\right] \\
&= \rho^2 \left[\frac{k(k-1)^2}{16} - 1\right] \\
&\geq \frac{(\rho k)^2}{32}.
\end{aligned}
$$

Now assume that $\rho k > 1$, then for $k > 32$,

$$
\begin{aligned}
\mathrm{KL}&(\mathbb{Q}_0 \,||\, \mathbb{Q}_1) \\
&\geq \frac{\rho^2 k}{16} - \frac{1}{2}\left[1 + 2\rho^2\right] \\
&\quad + \frac{1}{2}\left[\frac{1}{1+(k-1)\rho} + \log(1+(k-1)\rho)\right] \\
&\geq \rho^2\left[\frac{k}{16} - 1\right] \\
&\quad + \frac{1}{2}\left[\frac{1}{1+(k-1)\rho} + \log(1+(k-1)\rho) - 1\right] \\
&\geq \frac{\rho^2 k}{32} + \frac{\log(1+(k-1)\rho) - 1}{2}. \qquad \square
\end{aligned}
$$

APPENDIX B
EXTENSIONS TO UNNORMALIZED MODEL

A. Uniform (Non-Adaptive) Lower Bound for Detection of Positive Correlations

Proposition 8: For any class $\mathcal{C}$, any $\rho \in [0, 0.9)$, the minimum risk in the normalized model (resp. the unnormalized model) under uniform (non-adaptive) sensing is bounded as

$$
R^* \geq \frac{1}{2} - \frac{1}{4}\sqrt{E\left[\cosh^m\left(\frac{8\rho Z}{1-\rho}\right)\right] - 1}
$$

$$
\bar{R}^* \geq \frac{1}{2} - \frac{1}{4}\sqrt{E\left[\cosh^m(8\rho Z)\right] - 1}
$$

where Z is the size of the intersection of two elements of $\mathcal{C}$ drawn independently and uniformly at random.

Proof: This is essentially a reproduction of the proof of [4] with minor modifications. The details are omitted. $\square$

B. Uniform (Non-Adaptive) Upper Bound for Detection of Positive Correlations

Let $H(b) = b - 1 - \log b$ for $b > 1$.

Proposition 9: Under uniform (non-adaptive) sensing, the localized square-sum test that rejects when

$$
Y_{scan} = \max_{S \in \mathcal{C}} \sum_{t=1}^m \left(\sum_{i \in S} X_i^t\right)^2
$$

exceeds

$$
\frac{1}{2}\left(\rho k^2 m + H^{-1}(3\log|\mathcal{C}|/m) - 1)km\right)
$$

is asymptotically powerful when

$$
\rho k \geq c_1 \max\left(\sqrt{\frac{\log|\mathcal{C}|}{m}}, \frac{\log|\mathcal{C}|}{m}\right)
$$

both for the normalized and unnormalized models.

Proof: This is proved in [4] for the normalized model. In the case of the unnormalized model, the test statistic is distributed as $k\chi_m^2$ under the null, and as $(k(1+\rho) + \rho k(k-1))\chi_m^2$ under the alternative, which changes only mildly the proof with respect to the normalized model. $\square$

C. KL Divergences

Proposition 10: We have

$$
\mathrm{KL}(\bar{\mathbb{P}}_0 \,||\, \bar{\mathbb{P}}_S) = \frac{\mathbf{1}_{k \geq 2}}{2}\left[-1 + \frac{1}{1+\rho k} + \log(1+\rho k)\right]. \quad (13)
$$

Proof: The KL divergence between $\bar{\mathbb{P}}_0$ and $\bar{\mathbb{P}}_S$ can be computed using the standard formula for KL divergence between two centered Gaussian vectors, with covariances matrices

$$
\Sigma_0 = I_n, \quad \Sigma_1 = \bar{\Sigma}_S.
$$

When $k = 0$, the divergence is zero, and we will thus assume $k \geq 1$. Up to a simultaneous permutation of rows and columns,

$$
\bar{\Sigma}_S = \begin{bmatrix} I_{n-k} & \\ & I_k + K_\rho(k) \end{bmatrix}
$$

where $K_\rho(k) \in \mathbb{R}^{k \times k}$ has coefficients equal to ρ everywhere. Like previously, $I_k + K_\rho(k)$ is diagonalizable, and has eigenvalue 1 with multiplicity $k - 1$, and eigenvalue $1 + \rho k$ with multiplicity one. As a consequence, for $k \geq 1$, we have

$$
\begin{aligned}
\log \det \bar{\Sigma}_S &= \log(1 + \rho k) \\
\mathrm{trace}\, \bar{\Sigma}_S^{-1} &= (n-1) + \frac{1}{1+\rho k}.
\end{aligned}
$$

This leads to

$$
\begin{aligned}
\mathrm{KL}(\bar{\mathbb{P}}_0 \,||\, \bar{\mathbb{P}}_S) &= \frac{1}{2}\left[\mathrm{trace}(\Sigma_1^{-1}\Sigma_0) - n - \log(\det\Sigma_0/\det\Sigma_1)\right] \\
&= \frac{1}{2}\left[(n-1) - n + \frac{1}{1+\rho k} + \log(1+\rho k)\right].
\end{aligned}
$$

$\square$

Proposition 11: For any $A \subset [n]$,

$$
\mathrm{KL}(\bar{\mathbb{P}}_0|_A \,||\, \bar{\mathbb{P}}_S|_A) \leq \min\left[\frac{\rho}{2}, \frac{\rho^2 k}{2}\right] |A \cap S|.
$$

Proof: First note since the KL divergences are independent of n, it is sufficient to use the expressions of Proposition 7 with a contaminated set of size $s = |A \cap S|$. As previously, we assume $s \geq 1$, as the result is trivial otherwise. Consider the unnormalized model, with KL divergence given in (13). Using (3), we obtain

$$
\mathrm{KL}(\bar{\mathbb{P}}_0|_A \,||\, \bar{\mathbb{P}}_S|_A) = \mathrm{KL}(\bar{\mathbb{P}}_0 \,||\, \bar{\mathbb{P}}_{A \cap S}) \leq \frac{\rho s}{2}.
$$

Using (4) we obtain

$$
\mathrm{KL}(\bar{\mathbb{P}}_0|_A \,||\, \bar{\mathbb{P}}_S|_A) = \mathrm{KL}(\bar{\mathbb{P}}_0 \,||\, \bar{\mathbb{P}}_{A \cap S}) \leq \frac{\rho^2 s^2}{2} \leq \frac{\rho^2 sk}{2}.
$$

Combining these last two inequalities yields the desired result. $\square$

Proposition 12: Assume that ρ converges to zero. There exists numerical constants C_3 and C_4 such that, when either

$$\rho k \to \infty \quad \text{and} \quad m \log(1 + \rho k) \geq C_3 \log \log(n/k),$$

or

$$\rho k \to 0 \quad \text{and} \quad \rho k \sqrt{m} \geq C_4 \sqrt{\log \log(n/k)},$$

the sequential thresholding procedure for disjoint k-intervals has risk converging to zero.

Proof: For the unnormalized model, when $\rho k > 1$, using (8),

$$\mathrm{KL}(\bar{\mathbb{Q}}_0 \, || \, \bar{\mathbb{Q}}_1) \geq \frac{\log(1 + \rho k)}{10}.$$

When $\rho k < 1$, using (7),

$$\mathrm{KL}(\bar{\mathbb{Q}}_0 \, || \, \bar{\mathbb{Q}}_1) \geq \frac{(\rho k)^2}{16}. \qquad \square$$

Proposition 13: Assume that ρ converges to zero. There exists numerical constants C_5 and C_6 such that, when either

$$\rho k \to \infty \quad \text{and} \quad \rho k m \geq C_5 \log \log(n/k),$$

or

$$\rho k \to 0 \quad \text{and} \quad \rho k \sqrt{m} \geq C_6 \sqrt{\log \log(n/k)},$$

the modified sequential thresholding procedure for disjoint k-intervals has risk converging to zero.

Proof: For the unnormalized model with $p = \left\lceil \frac{1}{\rho} \right\rceil$, when $\rho k > 1$, we have $\left\lceil \frac{1}{\rho} \right\rceil < k + 1$, and as a consequence,

$$\mathrm{KL}(\bar{\mathbb{P}}_0^p \, || \, \bar{\mathbb{P}}_S^p) \geq \frac{\log 2 - 1/2}{2} \geq \frac{1}{11}. \qquad \square$$

REFERENCES

[1] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, "On combinatorial testing problems," *Ann. Statist.*, vol. 38, no. 5, pp. 3063–3092, 2010.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, Mar. 2002.

[3] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2008, pp. 2454–2458.

[4] E. Arias-Castro, S. Bubeck, and G. Lugosi. (2012). "Detecting positive correlations in a multivariate sample." [Online]. Available: http://arxiv.org/abs/1202.5536

[5] E. Arias-Castro, S. Bubeck, and G. Lugosi, "Detection of correlations," *Ann. Statist.*, vol. 40, no. 1, pp. 412–435, 2012.

[6] E. Arias-Castro, E. J. Candès, and M. A. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 472–481, Jan. 2013.

[7] E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni, "Searching for a trail of evidence in a maze," *Ann. Statist.*, vol. 36, no. 4, pp. 1726–1757, 2008.

[8] Y. Baraud, "Non-asymptotic minimax rates of testing in signal detection," *Bernoulli*, vol. 8, no. 5, pp. 577–606, 2002.

[9] B. M. Barber and J. D. Lyon, "Detecting long-run abnormal stock returns: The empirical power and specification of test statistics," *J. Financial Econ.*, vol. 43, no. 3, pp. 341–372, 1997.

[10] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *Ann. Statist.*, vol. 41, no. 4, pp. 1780–1815, 2013.

[11] C. Butucea and Y. I. Ingster. (2011). "Detection of a sparse submatrix of a high-dimensional noisy matrix." [Online]. Available: http://arxiv.org/abs/1109.0898

[12] T. Cai, Z. Ma, and Y. Wu. (2013). "Optimal estimation and rank detection for sparse spiked covariance matrices." [Online]. Available: http://arxiv.org/abs/1305.3235

[13] R. M. Castro, "Adaptim sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, vol. 20 no. 4, pp. 2217–2246, 2014. [Online]. Available: http:llprojecteuclidmo~euclid.bjl1411314458.

[14] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, 2009, Art. ID 15.

[15] Y. Chen and A. Krause, "Near-optimal batch mode active learning and adaptive submodular optimization," in *Proc. 30th ICML*, 2013, pp. 160–168.

[16] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Ann. Statist.*, vol. 32, no. 3, pp. 962–994, 2004.

[17] P. Hall and J. Jin, "Innovated higher criticism for detecting sparse signals in correlated noise," *Ann. Statist.*, vol. 38, no. 3, pp. 1686–1732, 2010.

[18] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, "Sequentially designed compressed sensing," in *Proc. IEEE Statist. Signal Process. Workshop (SSP)*, Aug. 2012, pp. 401–404.

[19] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Selective sampling for sparse signal recovery," in *Proc. 12th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2009, pp. 216–223.

[20] A. Hero and B. Rajaratnam, "Hub discovery in partial correlation graphs," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 6064–6078, Sep. 2012.

[21] S. A. Hofmeyr, S. Forrest, and A. Somayaji, "Intrusion detection using sequences of system calls," *J. Comput. Secur.*, vol. 6, no. 3, pp. 151–180, 1998.

[22] Y. I. Ingster, "Some problems of hypothesis testing leading to infinitely divisible distribution," *Math. Methods Statist.*, vol. 6, no. 1, pp. 47–69, 1997.

[23] Y. I. Ingster, C. Pouet, and A. B. Tsybakov, "Classification of sparse high-dimensional vectors," *Philosoph. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 367, no. 1906, pp. 4427–4448, Nov. 2009.

[24] D. Janakiram, V. Adi Mallikarjuna Reddy, and A. V. U. Phani Kumar, "Outlier detection in wireless sensor networks using Bayesian belief networks," in *Proc. 1st Int. Conf. Commun. Syst. Softw. Middleware (COMSWARE)*, 2006, pp. 1–6.

[25] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, vol. 29, no. 2, pp. 295–327, 2001.

[26] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *J. Amer. Statist. Assoc.*, vol. 104, no. 486, pp. 682–693, 2009.

[27] J. Jung, B. Krishnamurthy, and M. Rabinovich, "Flash crowds and denial of service attacks: Characterization and implications for CDNs and web sites," in *Proc. 11th Int. Conf. World Wide Web*, 2002, pp. 293–304.

[28] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari, "A space–time permutation scan statistic for disease outbreak detection," *PLoS Med.*, vol. 2, no. 3, p. e59, 2005.

[29] J. Lin, E. Keogh, A. Fu, and H. Van Herle, "Approximations to magic: Finding unusual medical time series," in *Proc. 18th IEEE Symp. Comput.-Based Med. Syst.*, Jun. 2005, pp. 329–334.

[30] M. Malloy and R. Nowak, "On the limits of sequential testing in high dimensions," in *Proc. Conf. Rec. 45th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 1245–1249.

[31] M. Malloy and R. Nowak, "Sequential analysis in high-dimensional multiple testing and sparse recovery," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul./Aug. 2011, pp. 2661–2665.

[32] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, "Inferring internet denial-of-service activity," *ACM Trans. Comput. Syst.*, vol. 24, no. 2, pp. 115–139, 2006.

[33] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 631–636.

[34] S. Shekhar, C.-T. Lu, and P. Zhang, "Detecting graph-based spatial outliers: Algorithms and applications (a summary of results)," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 371–376.

[35] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 5, pp. 631–645, May 2007.

[36] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 1015–1022.

[37] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003.

[38] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer-Verlag, 2009.

[39] T. Van Phuong, L. X. Hung, S. J. Cho, Y.-K. Lee, and S. Lee, "An anomaly detection algorithm for detecting attacks in wireless sensor networks," in *Intelligence and Security Informatics*. New York, NY, USA: Springer-Verlag, 2006, pp. 735–736.

[40] H. Wang, D. Zhang, and K. G. Shin, "Detecting SYN flooding attacks," in *Proc. IEEE 21st Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 3. Jun. 2002, pp. 1530–1539.

Rui M. Castro received a Licenciatura degree in aerospace engineering in 1998 from the Instituto Superior Técnico, Technical University of Lisbon, Portugal, and a Ph.D. degree in electrical and computer engineering from Rice University, Houston, Texas in 2008. Between 1998 and 2000 he was a researcher with the Communication Theory and Pattern Recognition Group, Institute of Telecommunications, Lisbon, and in 2002 he held a visiting researcher position at the Mathematics Research Center, Bell Laboratories Research. He was a postdoctoral fellow at the University of Wisconsin in 2007-2008, and between 2008 and 2010 he held an Assistant Professor position in the department of electrical engineering at Columbia University. He is currently an assistant professor in the Department of Mathematics and Computer Science of the Eindhoven University of Technology (TU/e) in the Netherlands. His broad research interests include non-parametric statistics, learning theory, statistical signal and image processing, network inference, and pattern recognition. Mr. Castro received a Rice University Graduate Fellowship in 2000 and a Graduate Student Mentor Award from the University of Wisconsin in 2008.

Gábor Lugosi graduated in electrical engineering at the Technical University of Budapest in 1987, and received his Ph.D. from the Hungarian Academy of Sciences in 1991. Since 1996, he has been at the Department of Economics, Pompeu Fabra University. In 2006 he became an ICREA research professor. His research interest includes learning theory, nonparametric statistics, inequalities in probability, random structures, and information theory.

Pierre-André Savalle received a degree in engineering and applied mathematics from Ecole Centrale Paris in 2011. He is currently a Ph.D. candidate at Ecole Centrale Paris and Ecole Normale Supérieure de Cachan.