Electronic Journal of Statistics
Vol. 6 (2012) 547–558
ISSN: 1935-7524
DOI: 10.1214/12-EJS686

Detecting a vector based on linear measurements

Ery Arias-Castro

University of California, San Diego
Department of Mathematics
La Jolla, CA 92093-0112
e-mail: eariasca@ucsd.edu

Abstract: We consider a situation where the state of a system is represented by a real-valued vector $\mathbf{x} \in \mathbb{R}^n$. Under normal circumstances, the vector $\mathbf{x}$ is zero, while an event manifests as non-zero entries in $\mathbf{x}$, possibly few. Our interest is in designing algorithms that can reliably detect events — i.e., test whether $\mathbf{x} = 0$ or $\mathbf{x} \neq 0$ — with the least amount of information. We place ourselves in a situation, now common in the signal processing literature, where information on $\mathbf{x}$ comes in the form of noisy linear measurements $y = \langle \mathbf{a}, \mathbf{x} \rangle + z$, where $\mathbf{a} \in \mathbb{R}^n$ has norm bounded by 1 and $z \in \mathcal{N}(0,1)$. We derive information bounds in an active learning setup and exhibit some simple near-optimal algorithms. In particular, our results show that the task of detection within this setting is at once much easier, simpler and different than the tasks of estimation and support recovery.

AMS 2000 subject classifications: Primary 62C20, 62G10, 62H15.
Keywords and phrases: Signal detection, compressed sensing, adaptive measurements, normal mean model, sparsity, high-dimensional data.

Received December 2011.

1. Introduction

We consider a situation where the state of a system is represented by a real-valued vector $\mathbf{x} \in \mathbb{R}^n$. Under normal circumstances, the vector $\mathbf{x}$ is zero, while an event manifests as non-zero entries in $\mathbf{x}$, possibly few. Our interest is in the design of algorithms that reliably detect events — i.e., test whether $\mathbf{x} = 0$ or $\mathbf{x} \neq 0$ — with the least amount of information. We assume that we may learn about $\mathbf{x}$ via noisy linear measurements of the form

$$y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle + z_i, \tag{1}$$

where the measurement vectors $\mathbf{a}_i$'s have Euclidean norm bounded by 1 and the noise z_i's are i.i.d. standard normal. Assuming that we may take a limited number of linear measurements, the engineering is in choosing them in order to minimize the false alarm and missed detection rates. We derive information bounds, establishing some fundamental detection limits relating the signal strength and the number of linear measurements. The bounds we obtain apply to all adaptive schemes, where we may choose the ith measurement vector $\mathbf{a}_i$ based on the past measurements, i.e., we may choose $\mathbf{a}_i$ as a function of $(\mathbf{a}_1, y_1, \ldots, \mathbf{a}_{i-1}, y_{i-1})$.

1.1. Related work

Learning as much as possible about a vector based on a few linear measurements is one of the central themes of compressive sensing (CS) [4, 5, 8]. Most of this literature, as it relates to signal processing, has focused on the tasks of estimation and support recovery. Particularly in surveillance situations, however, it makes sense to perform detection before estimation because, as we shall confirm, reliable detection is possible at much lower signal-to-noise ratios or, equivalently, with much fewer linear measurements than estimation. This can be achieved with much greater implementation ease and much lower computational cost than standard CS methods based on convex programming.

The literature on the detection of a high-dimensional signal is centered around the classical normal mean model, based on observations $y_i = x_i + z_i$, where the z_i's are i.i.d. standard normal. In this model, only one noisy observation is available per coordinate, so that some assumptions are necessary and the most common one, by far, is that the vector $\mathbf{x} = (x_1, \dots, x_n)$ is sparse. This setting has attracted a fair amount attention [7, 15, 16], with recent publications allowing adaptive measurements [12]. More recently, a few papers [2, 10, 14] extended these results to testing for a sparse coefficient vector in a linear system with the aim of characterizing the detection feasibility. These papers work with designs having low mutual coherence, for example, assuming that the $\mathbf{a}_i$'s are i.i.d. multivariate normal. As we shall see below, such designs are not always desirable. We also mention [13], which assumes that an estimator $\widehat{\mathbf{x}}$ of $\mathbf{x}$ is available and examines the performance of the test based on $\langle \widehat{\mathbf{x}}, \mathbf{x} \rangle$; and [17], which proposes a Bayesian approach for the detection of sparse signals in a sensor network for which the design matrix is assumed to have some polynomial decay in terms of the distance between sensors.

We mention that the present paper may be seen as a companion paper to [1] which considers the tasks of estimation and support recovery in the same setting.

1.2. Notation and terminology

Our detection problem translates into a hypothesis testing problem $H_0 : \mathbf{x} = 0$ versus $H_1 : \mathbf{x} \in \mathcal{X}$, for some subset $\mathcal{X} \subset \mathbb{R}^n \setminus \{0\}$. A test procedure based on m measurements of the form (1) is a binary function of the data, i.e., $T = T(\mathbf{a}_1, y_1, \dots, \mathbf{a}_m, y_m)$, with $T = \varepsilon \in \{0, 1\}$ indicating that T favors H_ε. The (worst-case) risk of a test T is defined as

$$\gamma(T) := \mathbb{P}_0(T = 1) + \max_{\mathbf{x} \in \mathcal{X}} \mathbb{P}_{\mathbf{x}}(T = 0),$$

where $\mathbb{P}_{\mathbf{x}}$ denotes the distribution of the data when $\mathbf{x}$ is the true underlying vector. With a prior π on the set of alternatives $\mathcal{X}$, the corresponding average (Bayes) risk is defined as

$$\gamma_\pi(T) := \mathbb{P}_0(T = 1) + \mathbb{E}_\pi \, \mathbb{P}_{\mathbf{x}}(T = 0),$$

where $\mathbb{E}_\pi$ denotes the expectation under π. Note that for any prior π and any test procedure T,

$$\gamma(T) \geq \gamma_\pi(T). \tag{2}$$

For a vector $\mathbf{a} = (a_1, \dots, a_k)$,

$$\|\mathbf{a}\| = \Big(\sum_j a_j^2\Big)^{1/2}, \qquad |\mathbf{a}| = \sum_j |a_j|,$$

and $\mathbf{a}^T$ denote its transpose. For a matrix $\mathbf{M}$,

$$\|\mathbf{M}\|_{\text{op}} = \sup_{\mathbf{a} \neq 0} \frac{\|\mathbf{M}\mathbf{a}\|}{\|\mathbf{a}\|}.$$

Everywhere in the paper, $\mathbf{x} = (x_1, \dots, x_n)$ denotes the unknown vector, while $\mathbf{1}$ denotes the vector with all coordinates equal to 1 and dimension implicitly given by the context.

1.3. Content

In Section 2 we focus on vectors $\mathbf{x}$ with non-negative coordinates. This situation leads to an exceedingly simple, yet near-optimal procedure based on a measurement scheme that is completely at odds with what is commonly used in CS. In Section 3 we treat the case of a general vector $\mathbf{x}$ and derive another simple, near-optimal procedure. In both cases, the methods we suggest are non-adaptive — in the sense that the measurement vectors are chosen independently of the observations — yet perform nearly as well as any adaptive method. In Section 4 we discuss our results and important extensions, particularly to the case of structured signals.

2. Vectors with non-negative entries

Vectors with non-negative entries may be relevant in image processing, for example, where the object to be detected is darker (or lighter) than the background. As we shall see, detecting such a vector is essentially straightforward in every respect. In particular, the use of low-coherence designs is counter-productive in this situation.

The first thing that comes to mind, perhaps, is gathering strength across coordinates by measuring $\mathbf{x}$ with the constant vector $\mathbf{1}/\sqrt{n}$. And, with a budget of m measurements, we simply take this measurement m times.

Proposition 1. *Suppose we take m measurements of the form (1) with $\mathbf{a}_i = \mathbf{1}/\sqrt{n}$ for all i. Consider then the test that rejects when*

$$\sum_{i=1}^m y_i > \tau\sqrt{m},$$

where τ is some critical value. Its risk against a vector $\mathbf{x}$ is equal to

$$1 - \Phi(\tau) + \Phi(\tau - \sqrt{m/n}|\mathbf{x}|),$$

where Φ is the standard normal distribution function. Hence, if $\tau = \tau_n \to \infty$, this test has vanishing risk against alternatives satisfying $\sqrt{m/n}|\mathbf{x}| - \tau_n \to \infty$.

Since we may chose $\tau_m \to \infty$ as slowly as we wish, in essence, the simple sum test based on repeated measurements from the constant vector has vanishing risk against alternatives satisfying $\sqrt{m/n}|\mathbf{x}| \to \infty$.

Proof. The result is a simple consequence of the fact that

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} y_i \sim \mathcal{N}(\sqrt{m/n}|\mathbf{x}|, 1).$$

□

Although the choice of measurement vectors and the test itself are both exceedingly simple, the resulting procedure comes close to achieving the best possible performance in this particular setting, as the following information bound reveals.

Theorem 1. *Let $\mathcal{X}(\mu, S)$ denote the set of vectors in $\mathbb{R}^n$ having exactly S non-zero entries all equal to $\mu > 0$. Based on m measurements of the form (1), possibly adaptive, any test for $H_0 : \mathbf{x} = 0$ versus $H_1 : \mathbf{x} \in \mathcal{X}(\mu, S)$ has risk at least $1 - \sqrt{m/(8n)}S\mu$.*

In particular, the risk against alternatives $\mathbf{x} \in \mathcal{X}(\mu, S)$ with $\sqrt{m/n}|\mathbf{x}| = \sqrt{m/n}S\mu \to 0$, goes to 1 uniformly over all procedures.

Proof. The standard approach to deriving uniform lower bounds on the risk is to put a prior on the set of alternatives and use (2). We simply choose the uniform prior on $\mathcal{X}(\mu, S)$, which we denote by π. The hypothesis testing problem reduces to $H_0 : \mathbf{x} = 0$ versus $H_1 : \mathbf{x} \sim \pi$, for which the likelihood ratio test is optimal by the Neyman-Pearson fundamental lemma. The likelihood ratio is defined

$$L := \frac{\mathbb{P}_\pi(\mathbf{a}_1, y_1, \dots, \mathbf{a}_m, y_m)}{\mathbb{P}_0(\mathbf{a}_1, y_1, \dots, \mathbf{a}_m, y_m)} = \mathbb{E}_\pi \exp\left(\sum_{i=1}^{m} y_i(\mathbf{a}_i^T \mathbf{x}) - (\mathbf{a}_i^T \mathbf{x})^2/2 \right),$$

where $\mathbb{E}_\pi$ denotes the expectation with respect to π, and the related test is $T = \{L > 1\}$. It has risk equal to

$$\gamma_\pi(T) = 1 - \frac{1}{2}\|\mathbb{P}_\pi - \mathbb{P}_0\|_{\mathrm{TV}}, \tag{3}$$

where $\mathbb{P}_\pi := \mathbb{E}_\pi \mathbb{P}_{\mathbf{x}}$ (the π-mixture of $\mathbb{P}_{\mathbf{x}}$) and $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance [18, Th. 2.2]. By Pinsker's inequality [18, Lem. 2.5]

$$\|\mathbb{P}_\pi - \mathbb{P}_0\|_{\mathrm{TV}} \le \sqrt{K(\mathbb{P}_0, \mathbb{P}_\pi)/2}, \tag{4}$$

where $K(\mathbb{P}_0, \mathbb{P}_\pi)$ denotes the Kullback-Leibler divergence [18, Def. 2.5]. We have

$$\begin{aligned} K(\mathbb{P}_0, \mathbb{P}_\pi) &= -\mathbb{E}_0 \log L && (5) \\ &\leq -\mathbb{E}_\pi \sum_{i=1}^m \mathbb{E}_0 \left(y_i(\mathbf{a}_i^T \mathbf{x}) - (\mathbf{a}_i^T \mathbf{x})^2/2 \right) && (6) \\ &= \mathbb{E}_\pi \sum_{i=1}^m \mathbb{E}_0 (\mathbf{a}_i^T \mathbf{x})^2/2 && (7) \\ &= \sum_{i=1}^m \mathbb{E}_0 \left(\mathbf{a}_i^T \mathbf{C} \mathbf{a}_i \right) && (8) \\ &\leq m \|\mathbf{C}\|_{\mathrm{op}}, && (9) \end{aligned}$$

where $\mathbf{C} = (c_{jk}) := \mathbb{E}_\pi(\mathbf{x}\mathbf{x}^T)$. The first line is by definition; the second is by definition of $\mathbb{P}_\mathbf{x} / \mathbb{P}_0$, by the application of Jensen's inequality justified by the convexity of $x \to -\log x$, and by Fubini's theorem; the third is by independence of $\mathbf{a}_i$, y_i and $\mathbf{x}$ (under $\mathbb{P}_0$), and by the fact that $\mathbb{E}(y_i) = 0$; the fourth is by independence of $\mathbf{a}_i$ and $\mathbf{x}$ (under $\mathbb{P}_0$) and by Fubini's theorem; the fifth is because $\|\mathbf{a}_i\| \leq 1$ for all i.

Since under π the support of $\mathbf{x}$ is chosen uniformly at random among subsets of size S, we have

$$c_{jj} = \mu^2 \, \mathbb{P}_\pi(x_j \neq 0) = \mu^2 \cdot \frac{S}{n}, \quad \forall j,$$

and

$$c_{jk} = \mu^2 \, \mathbb{P}_\pi(x_j \neq 0, x_k \neq 0) = \mu^2 \cdot \frac{S}{n} \cdot \frac{S-1}{n-1}, \qquad j \neq k.$$

This simple matrix has operator norm $\|\mathbf{C}\|_{\mathrm{op}} = \mu^2 S^2/n$.

Coming back to the divergence, we therefore have

$$K(\mathbb{P}_0, \mathbb{P}_\pi) \leq m \cdot \mu^2 S^2/n,$$

and returning to (3) via (4), we bound the risk of the likelihood ratio test as follows

$$\gamma(T) \geq 1 - \sqrt{K(\mathbb{P}_0, \mathbb{P}_\pi)/8} \geq 1 - \sqrt{m/(8n)} S\mu.$$

□

With Proposition 1 and Theorem 1, we conclude that the following is true in a minimax sense:

> *Reliable detection of a nonnegative vector $\mathbf{x} \in \mathbb{R}^n$ from m noisy linear measurements is possible if $\sqrt{m/n}|\mathbf{x}| \to \infty$ and impossible if $\sqrt{m/n}|\mathbf{x}| \to 0$.*

3. General vectors

When dealing with arbitrary vectors, the measurement vector $\mathbf{1}/\sqrt{n}$ may not be appropriate. In fact, the resulting procedure is completely insensitive to vectors

$\mathbf{x}$ such that $\langle \mathbf{1}, \mathbf{x}\rangle = 0$. Nevertheless, if one selects a measurement vector $\mathbf{a}$ from the Bernoulli ensemble — i.e., with independent entries taking values $\pm 1/\sqrt{n}$ with equal probability — then on average, $\langle \mathbf{a}, \mathbf{x}\rangle$ is of the order of $\|\mathbf{x}\|/\sqrt{n}$. This is true when the number of non-zero entries in $\mathbf{x}$ grows with the dimension n; if we repeat the process a few times, it becomes true for any fixed vector $\mathbf{x}$.

Proposition 2. *Sample $\mathbf{b}_1, \dots, \mathbf{b}_h$ independently from the Bernoulli ensemble, with $h \to \infty$ slowly, and take m measurements of the form (1) with $\mathbf{a}_i = \mathbf{b}_s$ for $i \in I_s := [(m/h)(s-1)+1, (m/h)s)$, $s = 1, \dots, h$. Consider then the test that rejects when*

$$\sum_{s=1}^{h} \left(\sum_{i \in I_s} y_i \right)^2 > 2m. \tag{10}$$

Its risk against a vector $\mathbf{x}$ satisfying $(m/n)\|\mathbf{x}\|^2 \geq 2h$ — averaged over the Bernoulli ensemble — is bounded from above by $42/h$. In particular, if $h = h_m \to \infty$, this test has vanishing risk against such alternatives.

Since we may take h_m increasing as slowly as we please, in essence, the test is reliable when $(m/n)\|\mathbf{x}\|^2 \to \infty$. Compared with repeatedly measuring with the constant vector $\mathbf{1}/\sqrt{n}$ as studied in Proposition 1, there is a substantial loss in power when $|\mathbf{x}|^2$ is much larger than $\|\mathbf{x}\|^2$. For example, when $\mathbf{x}$ has S non-zero entries all equal to $\mu > 0$, $|\mathbf{x}|^2 = S\|\mathbf{x}\|^2$.

Proof. For simplicity, assume that m/h is an integer and fix $\mathbf{x}$ throughout. For short, let

$$Y_s = \sum_{i \in I_s} y_i = (m/h)\langle \mathbf{b}_s, \mathbf{x}\rangle + \sqrt{m/h} Z_s, \quad Z_s := \sqrt{h/m} \sum_{i \in I_s} z_i.$$

Note that the Z_s's are i.i.d.$\sim \mathcal{N}(0,1)$, while the $\langle \mathbf{b}_s, \mathbf{x}\rangle$'s are i.i.d. with mean zero, variance $\|\mathbf{x}\|^2/n$ and fourth moment bounded by $6\|\mathbf{x}\|^4/n^2$ — which is immediate using the fact that the coordinates of $\mathbf{b}_s$ are i.i.d. taking values $\pm 1/\sqrt{n}$ with equal probability. Proceeding in an elementary way, we have

$$\begin{aligned} \mathbb{E}\left(\sum_{s=1}^{h} Y_s^2 \right) &= (m^2/h)\mathbb{E}\left(\langle \mathbf{b}_1, \mathbf{x}\rangle^2\right) + m\mathbb{E}\left(Z_1^2\right) \\ &= (m^2/h)\|\mathbf{x}\|^2/n + m, \end{aligned}$$

and

$$\begin{aligned} &\mathrm{Var}\left(\sum_{s=1}^{h} Y_s^2 \right) \\ = \; & (m^4/h^3)\, \mathbb{E}\left(\langle \mathbf{b}_1, \mathbf{x}\rangle^4\right) + 6(m^3/h^2)\mathbb{E}\left(\langle \mathbf{b}_1, \mathbf{x}\rangle^2\right)\mathbb{E}\left(Z_1^2\right) + (m^2/h)\mathbb{E}\left(Z_1^4\right) \\ \leq \; & 6(m^4/h^3)\|\mathbf{x}\|^4/n^2 + 6(m^3/h^2)\|\mathbf{x}\|^2/n + 3(m^2/h). \end{aligned}$$

Therefore, by Chebyshev's inequality, the probability of (10) under the null is bounded from above by $3/h$. Similarly, the probability of (10) *not* happening under an alternative $\mathbf{x}$ satisfying $(m/n)\|\mathbf{x}\|^2 \geq 2\tau\sqrt{h}$ is bounded from above by

$$\begin{aligned} & \frac{6(m^4/h^3)\|\mathbf{x}\|^4/n^2 + 6(m^3/h^2)\|\mathbf{x}\|^2/n + 3(m^2/h)}{((m^2/h)\|\mathbf{x}\|^2/n - m)^2} \\ \leq \quad & \frac{6(m^2/h)\|\mathbf{x}\|^4/n^2}{(m\|\mathbf{x}\|^2/n)^2/4} + \frac{6(m^3/h^2)\|\mathbf{x}\|^2/n}{h(m\|\mathbf{x}\|^2/n)/2} + \frac{3h}{h^2} \\ = \quad & \frac{24}{h} + \frac{12}{h} + \frac{3}{h} = \frac{39}{h}. \end{aligned}$$

For this, we used the fact that

$$\begin{aligned} (m^2/h)\|\mathbf{x}\|^2/n - m &= (m/h)\big((m\|\mathbf{x}\|^2/n) - h\big) \\ &\geq (m/h)\max\big((m\|\mathbf{x}\|^2/n)/2, h\big). \end{aligned}$$

Hence, this test has risk bounded above by $3/h + 39/h = 42/h$. □

Again, this relatively simple procedure nearly achieves the best possible performance.

Theorem 2. *Let $\mathcal{X}^{\pm}(\mu, S)$ denote the set of vectors in $\mathbb{R}^n$ having exactly S non-zero entries all equal to $\pm\mu$. Based on m measurements of the form (1), possibly adaptive, any test for $H_0 : \mathbf{x} = 0$ versus $H_1 : \mathbf{x} \in \mathcal{X}^{\pm}(\mu, S)$ has risk at least $1 - \sqrt{Sm/(8n)}\mu$.*

In particular, the risk against alternatives $\mathbf{x} \in \mathcal{X}^{\pm}(\mu, S)$ with $(m/n)\|\mathbf{x}\|^2 = (m/n)S\mu^2 \to 0$, goes to 1 uniformly over all procedures.

Proof. Again, we choose the uniform prior on $\mathcal{X}^{\pm}(\mu, S)$. The proof is then completely parallel to that of Theorem 1, now with $\mathbf{C} = \mu^2(S/n)\mathbf{I}$ — since the signs of the nonzero entries of $\mathbf{x}$ are i.i.d. Rademacher — so that $\|\mathbf{C}\|_{\text{op}} = \mu^2 S/n$. □

With Proposition 2 and Theorem 2, we conclude that the following is true in a minimax sense:

> *Reliable detection of a vector $\mathbf{x} \in \mathbb{R}^n$ from m noisy linear measurements is possible if $\sqrt{m/n}\|\mathbf{x}\| \to \infty$ and impossible if $\sqrt{m/n}\|\mathbf{x}\| \to 0$.*

4. Discussion

In this short paper, we tried to convey some very basic principles about detecting a high-dimensional vector with as few linear measurements as possible. First, when the vector has non-negative entries, repeatedly sampling from the constant vector $\mathbf{1}/\sqrt{n}$ is near-optimal. Second, when the vector is general but sparse, repeatedly sampling from a few measuring vectors drawn from a standard random (e.g., Bernoulli) ensemble is also near-optimal. In both cases, choosing the measuring vectors adaptively does not bring a substantial improvement. And, moreover, sparsity does not help, in the sense that the detection rates depend on $|\mathbf{x}|$ and $\|\mathbf{x}\|$, respectively.

4.1. A more general adaptive scheme

Suppose we may take as many linear measurements of the form (1) as we please (possibly an infinite number), with the only constraint being on the total measurement energy

$$\sum_i \|\mathbf{a}_i\|^2 \le m. \tag{11}$$

(Note that m is no longer constrained to be an integer.) This is essentially the setting considered in [11, 12], and clearly, the setup we studied in the previous sections satisfies this condition. So what can we achieve with this additional flexibility?

In fact, the same results apply. The lower bounds in Theorem 1 and Theorem 2 are proved in exactly the same way. (We effectively use (11) to go from (8) to (9), and this is the only place where the constraints on the number and norm of the measurement vectors are used.) Of course, Proposition 1 and Proposition 2 apply since the measurement schemes used there satisfy (11). However, in this special case they could be simplified. For instance, in Proposition 1 we could take one measurement with the constant vector $\sqrt{m/n}\,\mathbf{1}$.

4.2. Detecting structured signals

The results we derived are tailored to the case where $\mathbf{x}$ has no known structure. What if we know a priori that the signal $\mathbf{x}$ has some given structure? The most emblematic case is when the support of $\mathbf{x}$ is an interval of length S. In the classical setting where each coordinate of $\mathbf{x}$ is observed once, the scan statistic (aka generalized likelihood ratio test) is the tool of choice [3]. How does the story change in the setting where adaptive linear measurements in the form of (1) can be taken?

Perhaps surprisingly, knowing that $\mathbf{x}$ has such a specific structure does not help much. Indeed, Theorem 1 and Theorem 2 are proved in the same way. In the case of non-negative vectors, we use the uniform prior on vectors with support an interval of length S and nonzero entries all equal to μ, and the proof is identical, except for the matrix $\mathbf{C}$, which now has coefficients $c_{jk} = \mu^2 \max(S-|j-k|, 0)/n$ for all j, k. Because $\mathbf{C}$ is symmetric, we have

$$\|\mathbf{C}\|_{\mathrm{op}} \le \max_j \sum_k |c_{jk}| = \mu^2 S^2/n, \tag{12}$$

which is exactly the same bound as before. In the general case, the arguments are really identical, except that we use the uniform prior on vectors with support an interval of length S and nonzero entries all equal to μ in absolute value. (Here the matrix $\mathbf{C}$ is exactly the same.) Of course, Proposition 1 and Proposition 2 apply here too, so the conclusions are the same. Also, these conclusions hold in the more general setup with measurements satisfying (11).

To appreciate how powerful the ability to take linear measurements in the form of (1) with the constraint (11) really is, let us stay with the same task of detecting an interval of length S with a positive mean. On the one hand, we have the simple test based on $\sum_i y_i$ studied in Proposition 1. On the other hand, we have the scan statistic

$$\max_t \sum_{i=t}^{t+S-1} y_i,$$

with observations of the form

$$y_i = x_i + \sigma z_i, \quad \sigma := \sqrt{n/m}. \tag{13}$$

While the former requires $\sqrt{m/n}|\mathbf{x}| \to \infty$ to be asymptotically powerful, the scan statistic requires

$$\liminf \sqrt{m/n}|\mathbf{x}| \cdot (S \log^+(n/S))^{-1/2} \geq \sqrt{2},$$

where $\log^+(x) := \max(\log x, 1)$. With observations provided in the form of (13), this is asymptotically optimal [3]. Note that (13) is a special case of (11). Hence, the ability to take measurements of the form (1) allows to detect structured signals that are potentially much weaker, without a priori knowledge of the structure and with much simpler algorithms. Hardware that is able to take linear measurements such as (1) is currently being developed [9].

4.3. A comparison with estimation and support recovery

The results we obtain for detection are in sharp contrast with the corresponding results in estimation and support recovery. Though, by definition, detection is always easier, in most other settings it is not that much easier. For example, take the normal mean model described in the Introduction, assuming $\mathbf{x}$ is sparse with S coefficients equal to $\mu > 0$. In the regime where $S = n^{1-\beta}$, $\beta \in (1/2, 1)$, detection is impossible when $\mu \leq \sqrt{2r \log n}$ with $r < \rho_1(\beta)$, while support recovery is possible when $\mu \geq \sqrt{2r \log n}$ with $r > \rho_2(\beta)$, for a fixed functions $\rho_1, \rho_2 : (1/2, 1) \to (0, \infty)$ [7, 15, 16]. So the difference is a constant factor in the per-coordinate amplitude. In the setting we consider here, we are able to detect at a much smaller signal-to-noise ratio than what is required for estimation or support recovery, which nominally require at least $m \geq S$ measurements regardless of the signal amplitude, where S is the number of nonzero entries in $\mathbf{x}$. In detection, however, we saw that $m = 1$ measurement may suffice if the signal amplitude is large enough. Also, [1] shows that reliable support recovery is impossible when $\mu\sqrt{m/n} \to 0$, while we saw that $\mu S\sqrt{m/n} \to \infty$ and $\mu\sqrt{Sm/n} \to \infty$ suffice for reliable detection in the nonnegative and general cases, respectively. Therefore, having the ability to take linear measurements of the form (1) in a surveillance setting, it makes sense to perform detection as described here before estimation (identification) or support recovery (localization) of the signal.

4.4. Possible improvements

Though we provided simple algorithms that nearly match information bounds, there might be room for improvement. For one thing, it might be possible to reliably detect when, say, $\sqrt{m/n}|\mathbf{x}|$ is sufficiently large — for the case where $x_j \geq 0$ for all j — without necessarily tending to infinity. A good candidate for this might be the Bayesian algorithm proposed in [6].

More importantly, in the general case of Section 3, we might want to design an algorithm that detects any fixed $\mathbf{x}$ with high-probability, without averaging over the measurement design. This averaging may be interpreted in at least two ways:

(A1) If we were to repeat the experiment many times, each time choosing new measurement vectors and corrupting the measurements with new noise, then for a fixed vector $\mathbf{x}$, in most instances the test would be accurate.

(A2) Given the amplitudes $|x_j|,\ j = 1, \dots, n$, for most sign configurations the test will be accurate.

Interpretation (A1) is controversial as we do not repeat the experiment, which would amount to taking more samples. And interpretation (A2) raises the issue of robustness to any sign configuration. One way — and the only way we know of — to ensure this robustness is to use a CS-like sampling scheme, i.e., choosing $\mathbf{a}_1, \dots, \mathbf{a}_m$ in (1) such that the matrix with these rows satisfies RIP-like properties. This setting is studied in detail in [2], which in a nutshell says the following. Take measurement vectors from the Bernoulli ensemble, say, but hold the measurement design fixed. This is just a way to build a measurement matrix satisfying the RIP and with low mutual coherence. In particular, this requires that m is of order at least $S \log n$, though what follows assumes that $m \gg S(\log n)^3$. Based on such measurements, the test based on $\sum_i y_i^2$ is able to detect when $(\sqrt{m}/n)\|\mathbf{x}\|^2 \to \infty$, which is more stringent than what is required in Proposition 2; while the test based on $\max_{j=1,\dots,n} |\sum_i a_{ij}\, y_i|$ is able to detect when $\liminf \sqrt{m/n} \max_j |x_j|(\log n)^{-1/2} > \sqrt{2}$, which, except for the log factor, is what is required for support recovery. And this is essentially optimal, as shown in [2].

Acknowledgements

The author would like to thank Emmanuel Candès for stimulating discussions, Rui Castro for clarifying the set up described in Section 4.1 and two anonymous referees for their careful reading and constructive feedback. Some of the ideas presented here originated at the “New Trends in Mathematical Statistics” conference held at the Center International de Rencontres Mathématiques (CIRM) in Luminy, France, in 2011. This work was partially supported by a grant from the Office of Naval Research (N00014-09-1-0258).

References

[1] Arias-Castro, E., Candès, E. J. and Davenport, M. A. (2011). On the Fundamental Limits of Adaptive Sensing. Submitted.

[2] Arias-Castro, E., Candès, E. J. and Plan, Y. (2011). Global Testing under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism. *Ann. Statist.* **39** 2533–2556.

[3] Arias-Castro, E., Donoho, D. L. and Huo, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51** 2402–2425. MR2246369

[4] Candès, E. and Tao, T. (2006). Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Trans. Inform. Theory* **52** 5406–5425. MR2300700

[5] Candès, E. J., Romberg, J. and Tao, T. (2006). Robust uncertainty principles: Exact Signal Reconstruction from Highly Incomplete Fourier Information. *IEEE Trans. Info. Theory* **52** 489–509. MR2236170

[6] Castro, R. M., Haupt, J., Nowak, R. and Raz, G. M. (2008). Finding needles in noisy haystacks. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* 5133 -5136.

[7] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

[8] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

[9] Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F. and Baraniuk, R. G. (2008). Single-pixel imaging via compressive sampling. *Signal Processing Magazine, IEEE* **25** 83–91.

[10] Duarte, M. F., Davenport, M. A., Wakin, M. B. and Baraniuk, R. G. (2006). Sparse Signal Detection from Incoherent Projections. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* **3** III.

[11] Haupt, J., Baraniuk, R., Castro, R. and Nowak, R. (2009). Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements In *Proc. 43rd Asilomar Conf. on Signals, Systems, and Computers.*

[12] Haupt, J., Castro, R. and Nowak, R. (2009). Distilled sensing: Selective sampling for sparse signal recovery. In *Proc. 12th International Conference on Artificial Intelligence and Statistics (AISTATS)* 216–223. Citeseer.

[13] Haupt, J. and Nowak, R. (2007). Compressive sampling for signal detection. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).*

[14] Ingster, Y. I., Tsybakov, A. B. and Verzelen, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics* **4** 1476–1526. MR2747131

[15] Ingster, Y. I. (1999). Minimax detection of a signal for ℓ_n^l balls. *Math. Methods Statist.* **7** 401–428. MR1680087

[16] Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*. *Lecture Notes in Statistics* **169**. Springer-Verlag, New York. MR1991446

[17] Meng, J., Li, H. and Han, Z. (2009). Sparse Event Detection in Wireless Sensor Networks using Compressive Sensing. In *43rd Annual Conference on Information Sciences and Systems (CISS)*.

[18] Tsybakov, A. (2009). *Introduction to nonparametric estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359