Rui M. Castro
Department of Mathematics and Computer Science
Eindhoven University of Technology
rmcastro@tue.nl

Ervin Tánczos
Wisconsin Institute for Discovery
University of Wisconsin–Madison, USA
tanczos@wisc.edu

On adaptive sensing for high-dimensional signal inference

In many practical settings one can sequentially and adaptively guide the collection of future data, based on information extracted from data collected previously. These sequential data collection procedures are known by different names, such as sequential experimental design, active learning or adaptive sensing/sampling. In this paper Rui Castro and Ervin Tánczos give a small overview of recent results characterizing the fundamental limits of adaptive sensing for sparse signal inference. They focus primarily on support estimation for static signals, and detection of dynamically evolving signals. They show that, in many situations, adaptive sensing is able to make reliable inferences in situations where non-adaptive sensing would fail dramatically.

At large, the goal of any statistical learning and inference methodology is to uncover important aspects of a given process (e.g., physical, social or other) by collecting and analyzing data. Naturally, both the collection and analysis of data are essential aspects of the methodology, and as such should not be considered separately. In all but few cases one can view the data collection process as a 'querying' or experimenting procedure, where collected data are 'answers' or outcomes of multiple queries/experiments, frequently corrupted in some fashion.

Often there is significant flexibility in the choice of queries. If this choice must be made before *any* data is collected this corresponds to the classical experiment design scenario. However, if each query/experiment can be chosen sequentially, shaped by the answers/outcomes of the previous queries, much more powerful data collection and inference approaches can be used. Depending on the research area this sensing paradigm is known by different names: *sequential experimental design* in the statistics and economics literature, *active learning* [2,7,15] or *adaptive sensing/sampling* in computer science, engineering and machine learning and statistics. An essential aspect of adaptive sensing is the intricate coupling between data analysis and acquisition, which creates a powerful feedback structure. This is a double-edged sword: it is key to harness the power of sequential experimental design but also makes the analysis and design of these procedures rather challenging — indeed it creates complicated and strong dependencies in the data.

More broadly, the feedback between data analysis and acquisition is key in the scientific discovery method, a process of complex interactions between experiments and outcomes, guided mostly by the scientists intuition. There have been a few attempts to semi-automate such processes, notably the efforts reported in [17], where a large team of scientists developed a robot capable of autonomously conducting experiments and test hypothesis leading to the discovery of novel genomic knowledge in yeast cells.

The goal of this paper is to give a small overview of recent results pertaining inference of signals living in high dimensional spaces, as well as an idea of the methods and techniques used in the development of the analysis and algorithms. Although these results pertain only a certain class of settings and problems many of the surrounding methods and ideas are extremely useful in a broader sense.

A sparse signal model

In this paper we survey some recent results pertaining inference of signals *living* in high dimensional spaces. These signals might be static, meaning they do not change during the measurement period, or dynamic, meaning they evolve while data is being collected. For simplicity, we consider first static signals. Concretely, the signal of interest is represented by an n-dimensional vector $\mathbf{x} \in \mathbb{R}^n$, where n is called the *extrinsic signal dimension*. Depending on the context, the entries of $\mathbf{x}$ might be used to represent different quantities. For instance, in gene expression studies, each entry of the signal vector is associated with a different gene, and the value of the signal is related to the corresponding expression level. Another compelling example pertains the monitoring of computer networks for detection and localization of anomalous behavior. In that case each entry of $\mathbf{x}$ might represent the activity level of each node in the network.

Although in principle the vector $\mathbf{x}$ can be arbitrary, in most cases it is *sparse* meaning most entries of $\mathbf{x}$ have nominal/typical values, and only relatively few entries have values deviating from that norm. Referring back to the two examples above most genes will be expressed to their nominal value, and only a few genes (e.g., about 100 out of 20.000 genes) will have a different expression level, for instance, due to the onset of a disease. In the case of network monitoring most nodes in the network will behave normally, and only a few nodes will have higher activity, for instance, if those nodes are participating in a distributed denial-of-service attack.

To abstract this assumption let S be a subset of $\{1,\ldots,n\}$ of non-zero entries of $\mathbf{x}$ and assume that for all $i \in \{1,\ldots,n\}$ such that $i \notin S$ we have $x_i = 0$ (the nominal value of each entry). We refer to S as the *signal support* and this is our main object of interest. We might want to estimate the signal support set, or simply detect if S is not the empty set, as we formalize below. In the section 'Dynamically evolving signals' we consider also a modification of this model to allow signals to evolve over time.

To greatly simplify the presentation in this paper we consider only signals of the form

$$x_i = \begin{cases} \mu & \text{if } i \in S, \\ 0 & \text{if } i \notin S, \end{cases}$$

where $\mu > 0$ is called the *signal amplitude*. This restriction is also considered in [1,11] in the non-adaptive sensing context and does not substantially hinder the generality of the results presented in this manuscript. In particular, when characterizing the difficulty of the problem in terms of the minimum signal magnitude $\min_i |x_i|$ the characterization we provide is essentially sharp. Throughout this paper we consider the sparse regime where $|S| \ll n$ (meaning $|S|$ is much smaller than n). More specifically we assume $|S| \le n/\log_2(n)$.

Measurement model

Naturally, we do not have access to the signal $\mathbf{x}$ directly. Rather, we can collect only partial information through noisy measurements of individual entries, or possible ensembles of the signal entries. The latter is often referred to as Compressive Sensing (see [9,12] and references therein). In this paper we focus primarily on the first model. This sensing model was introduced in [16]. We assume it is possible to collect measurements of each signal component corrupted by additive Gaussian noise. Concretely let

$$Y_k = x_{A_k} + \Gamma_k^{-1/2} W_k, \quad k = 1, 2, \ldots,$$

where k denotes the measurement index (first measurement, second measurement, and so on), A_k denotes the entry of $\mathbf{x}$ being measured, Γ_k denotes the corresponding *precision* of the measurement, and W_k is a standard normal random variable embodying measurement noise. Importantly, W_k are independent of $\{Y_i\}_{i=1}^{k-1}$ and also independent of $\{A_i, \Gamma_i\}_{i=1}^{k}$. Said differently, each measurement corresponds to a single signal entry corrupted with additive Gaussian noise. Both choices of which entry to measure and the corresponding noise level can be determined by experimenter — the higher the precision of a measurement, the lower the noise level. Note also that it is possible to measure the same signal component multiple times, with independent noise realizations.

However, it is not possible to measure all the entries with arbitrarily large precision. In particularly, there is a total sensing budget constraint that must be satisfied, namely

$$\sum_{k=1}^{\infty} \mathbb{E}(\Gamma_k) \le m, \tag{1}$$

where $m > 0$. This means that it is not possible to measure all the signal entries with very high precision. Note that the above constraint is on the *expected* precision used. Alternatively, we can consider a slightly more stringent constraint on the actual precision (i.e., $\sum_{k=1}^{\infty} \Gamma_k \le m$). The formulation in equation (1) considerably simplifies the presentation, but does not qualitatively alter any of the results presented below. The reason is that, for most reasonable algorithms, control over the expected precision translates into control over the actual total precision in high probability by a concentration of measure argument.

This model might seem peculiar at first, as it allows for a scenario where one takes an infinite but countable number of measurements (provided the precision decays sufficiently fast as a function of k). Nevertheless, this model fits closely several practical situations, namely when the precision of measurements is proportional to the time it takes to collect a measurement. In that case equation (1) is stating one has to take measurements during a time period with the duration of m units. A representative example of such a setting is in astronomical surveys, where long exposure times are used to reduce the noise level. It is also possible to consider instead settings where there is little or no control over the precision (e.g., say $\Gamma_k \equiv \Gamma$ for all $k \in \{1,2,\ldots\}$). In that case the constraint in equation (1) simply states there is a maximum number of measurements that are allowed. This situation is closely related to what is encountered in stochastic multi-armed bandits settings [3].

Allowing for arbitrary precision values has several advantages: it is a more general measurement model, and allows for cleaner analytic results. Nevertheless, the entire methodology and analysis can still be done when considering fixed precision measurements and a constraint on the total number of measurements. When discussing inference of dynamically evolving signals in the section 'Dynamically evolving signals' we actually consider that situation instead.

The measurement model above is rather general, and allows for adaptive sensing approaches. Namely, one can adjust the way new measurements are collected based on measurements collected earlier. In other words, one can choose A_k, Γ_k as a function of the past experiments $\{Y_i, A_i, \Gamma_i\}_{i=1}^{k-1}$. Furthermore, this choice can also incorporate extra randomness, if desired. The collection of conditional distributions of A_k, Γ_k given $\{Y_i, A_i, \Gamma_i\}_{i=1}^{k-1}$ for all k is referred to as the *sensing strategy*. We can also consider more traditional non-adaptive sensing strategies. In that case the choice of sensing actions and corresponding precision must be made before collecting any data. Formally this means that $\{A_k, \Gamma_k\}_{k \in \mathbb{N}}$ is statistically independent from $\{Y_k\}_{k \in \mathbb{N}}$. Note that a non-adaptive design can still be random.

The case $m = n$ is of particular interest, allowing for a simple direct comparison between adaptive and non-adaptive sensing methodologies. When $m = n$ we allow on average one unit of precision per each of the signal entries. So, if there is no reason to give preference to any particular entry of $\mathbf{x}$, the natural optimal non-adaptive sensing strategy should simply measure each entry of $\mathbf{x}$ exactly once, with precision one. This corresponds to the well studied normal means model.

Although we are considering measurements with additive Gaussian noise, the methodologies used to study inference problems in this setting are easily extended to more general distributions (see for instance [20]). Actually, the setting described here is closely related to a class of problems known as *stochastic multi-armed bandits* (see [3] for an excellent survey). In North-American slang, a *one-armed bandit* is a casino slot-machine. In the stochastic multi-armed bandit setting one can imagine a row of n slot machines. Each machine is endowed with an unknown probability distribution F_i with mean x_i, $i \in \{1, \ldots, n\}$. At each turn k the experimenter can choose a machine A_k to play, and will observe a 'reward', nothing more than a sample of F_{A_k}. Depending on the setting the experimenter might have different goals. He/she might want to maximize the total reward — this leads to an exploration/exploitation trade-off. Or instead they might want to identify the best-paying machine, leading to the so called pure-exploration problem. The latter is intimately related to the problem of signal detection described above, which has essentially the same statistical complexity as locating a single non-zero signal component. Actually, in [13] a methodology for proving lower bounds for the best-arm problem was developed, which is essentially the same methodology developed earlier for the detection problem [5].

Support estimation

As stated earlier, we focus mainly on two classes of inference problems — support estimation and signal detection. We start by addressing the problem of support estimation, in which the goal is to identify the signal support S. In other words, one desires to construct a set estimator $\hat{S}$, based on the data $\{A_k, \Gamma_k, Y_k\}_{k=1}^{\infty}$ that is 'close' to the true signal support S. There are different ways to measure the closeness of these two sets. For concreteness we focus on number of errors, which is given by the cardinality of the symmetric set difference $\hat{S}\Delta S = (S \setminus \hat{S}) \cup (\hat{S} \setminus S)$.

Let us assume $S \in C$, where C is a given class of non-empty subsets of $\{1, \ldots, n\}$. For simplicity of presentation we assume that all the sets in C have the same cardinality s. A prototypical example is the class of all support sets with cardinality s (consisting of $\binom{n}{s}$ sets). The goal of support set estimation is therefore to construct a support set estimator $\hat{S} \equiv \hat{S}(\{A_k, \Gamma_k, Y_k\}_{k=1}^{\infty})$ such that worst case error

$$\max_{S \in C} \mathbb{E}_S[\hat{S}\Delta S]$$

is small. In the above $\mathbb{E}_S$ denotes the joint probability distribution of $\{A_i, \Gamma_i, Y_i\}_{i=1}^{\infty}$ for a given support set S.

For support estimation other metrics can be considered, such as $\mathbb{P}(\hat{S} \neq S)$ or False Discovery Rate (FDR) plus Non-Discovery Rate (NDR). The first is very similar to what is described above while the second metric is more lenient: we only try to make the number of errors *relative to* the size of $\hat{S}$ and S small. This means weaker signals can be reliably recovered. See for instance [5,16].

Non-adaptive sensing

Provided the class C of possible support sets is somewhat symmetric, there is no a priori reason to measure any given signal component in detriment to another component. The formal definition of symmetric classes is given below.

Definition 1. Let $S \in C$ be drawn uniformly at random. If $\mathbb{P}(i \in S)$ has the same value for all $i = 1, \ldots, n$, the class is said to be *symmetric*.

Equivalently, if $\sum_{S \in C} \mathbb{1}\{i \in S\}$ is not a function of i, the class is said to be symmetric. In the previous expression $\mathbb{1}$ denotes the usual indicator function.

For symmetric classes there is no reason to measure any signal component with more precision than any other. Therefore, if considering the non-adaptive sensing paradigm, the only reasonable sensing strategy is to use *uniform sensing*. This means that we distribute our sensing budget uniformly over all the signal components. Furthermore, collecting more than one measurement per signal entry is not necessary, due to statistical sufficiency. Therefore the optimal non-adaptive sensing strategy consists of n measurements (one per signal components), each with precision m/n. Said differently, we collect n independent measurements $Y_i \sim \mathcal{N}(x_i, n/m)$, $i \in \{1, \ldots, n\}$.

Since this is an estimation problem it is sensible to consider a maximum likelihood approach. Namely, construct $\hat{S}$ by choosing the support set S that maximizes the likelihood of the observations. It is not hard to show that this gives rise to the estimator

$$\hat{S}_{\text{non-adaptive}} = \underset{S \in C}{\operatorname{argmax}} \sum_{i \in S} Y_i. \qquad (2)$$

When considering the class of all support sets of cardinality s this methodology simply deems the s largest observations as the support estimate. Naturally, to ensure the expected number of errors is small the signal magnitude μ must be above the 'noise floor'. In particular for an arbitrary $\varepsilon > 0$, if $\mu \geq \sqrt{\frac{2n}{m}(1+\varepsilon)\log n}$ then it is guaranteed that

$$\max_{S \in C} \mathbb{E}_S[\hat{S}_{\text{non-adaptive}}\Delta S] \to 0,$$

as $n \to \infty$. Conversely, if $\mu < \sqrt{\frac{2n}{m}(1-\varepsilon)\log n}$ such a methodology will necessarily fail. In fact, one can show that if $\mu < \sqrt{\frac{cn}{m}\log\frac{n}{s}}$ for c slightly smaller than $\frac{1}{2}$, then no method whatsoever will be able to reliably estimate the support. More generally it turns out that under mild conditions the cardinality of the class C is the main bottleneck in regards estimation of the signal support. The following result characterizes the limits of *any* support estimation procedure.

Proposition 1. (Proposition 10 of [8]) *Suppose that C is symmetric, it only contains sets of size s and that $1+\sqrt{2} \leq (1-2\varepsilon)\log(|C|-1)$, where $|C|$ denotes the cardinality of C. If*

$$\mu \leq \sqrt{(1-2\varepsilon)\frac{n}{2sm}\log(|C|-1)},$$

then no non-adaptive procedure can satisfy

$$\mathbb{P}(\hat{S} \neq S) \leq \varepsilon.$$

Proposition 1 deals with a different error metric than the one we considered before. Note however that $\mathbb{P}(\hat{S} \neq S) \leq \mathbb{E}[|\hat{S}\Delta S|]$, hence the proposition also holds with $\mathbb{P}(\hat{S} \neq S)$ replaced by $\mathbb{E}[|\hat{S}\Delta S|]$ in the statement.

For the class of all sets of cardinality s we have $|C| = \binom{n}{s} \approx (n/s)^s$, giving rise to the claim made earlier. In a nutshell, in order for non-adaptive sensing to be successful in recovering a sparse signal the signal magnitude needs to scale roughly like $\sqrt{\frac{n}{m}\log n}$ (since when $s \ll n$, $\log\frac{n}{s}$ has the same scaling as $\log n$). If the class C has some structure (so the support sets are not arbitrary subsets of $\{1,\dots,n\}$) slightly better results can be obtained (see [8] and the subsection 'Structured support estimation').

Adaptive sensing

Having established that it is not possible to recover the support of signals with non-adaptive sensing when magnitude of the active components is roughly smaller than $\sqrt{\frac{n}{m}\log n}$, a natural question to ask is if it is possible to do better with adaptive sensing. As the reader might expect, the answer is affirmative, and we describe next a simple approach and analysis demonstrating this.

Consider a procedure in which we test each entry $x_i, i \in \{1,\dots,n\}$ independently to assess if these are zero or not. To perform the test, for entry i we take repeated measurements with the same precision Γ. If a negative measurement is collected we interpret that as evidence supporting the case that $x_i = 0$; hence we terminate the test and decide $i \notin \hat{S}$. On the other hand, if we measure a component T times without ever seeing a negative value, we interpret it as evidence supporting $x_i = \mu > 0$; we terminate the test and decide $i \in \hat{S}$.

Remarkably, this simple procedure performs very well — in fact almost as well as the best possible procedure — as we illustrate in the analysis below. Note that we still need to specify the parameters Γ and T. A good choice will become apparent from the analysis.

To evaluate the performance of this method we need to do two things. On one hand we need to assess the error that this method incurs. On the other hand, we need to ensure the precision budget of equation (1) is not exceeded. We start with the latter.

Since all measurements are made with the same precision Γ, our task is to count the total number of measurements the procedure makes in expectation. The expectation of the total number of measurements is simply the sum of the expected number of measurements for each individual test. Note that there are $(n-s)$ zero components in x. In this case, the probability that we observe a negative value is $\frac{1}{2}$. Since we stop measuring a component once we see a negative value, the expected number of measurements in such cases is 2. On the other hand, the test never performs more than T measurements for any coordinate. Hence the expected precision used by the test can be upper bounded as

$$\mathbb{E}\left(\sum_k \Gamma_k\right) \leq \Gamma(2(n-s) + sT). \qquad (3)$$

This can be used to determine the precision Γ of each measurement once we determined the adequate choice for T.

The next step is to control the number of errors. Since the procedure tests each component independently, this means we only need to understand the probability of error for a single test. Consider the test of component i, where $i \in \{1,\dots,n\}$. In case $x_i = 0$, an error is made if we observe T consecutive positive measurements. Since the probability of any one measurement being positive is $\frac{1}{2}$, the probability of making an error (meaning $i \in \hat{S}$) is simply 2^{-T}.

In case $x_i = \mu$, an error is made unless all measurements are positive. Let Φ denote the cumulative distribution function of a standard normal distribution (specifically, if $Z \sim \mathcal{N}(0,1)$ is a standard normal random variable, then $\Phi(z) = \mathbb{P}(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ for $z \in \mathbb{R}$). Denoting the (possible) measurements by $Y_{i,1},\dots,Y_{i,T}$ we see that the error probability is upper bounded by

$$\begin{aligned}
&\mathbb{P}(\exists j \in \{1,\dots,T\}: Y_{i,j} < 0) \\
&\quad \leq \sum_{j=1}^{T} \mathbb{P}(Y_{i,j} < 0) \\
&\quad = T\Phi(-\sqrt{\Gamma}\,\mu) \\
&\quad \leq \frac{T}{2}\exp(-\Gamma\mu^2/2),
\end{aligned}$$

where $Y_{i,j} \sim \mathcal{N}(\mu, 1/\Gamma)\ \forall j \in \{1,\dots,T\}$, and we used a union bound and a standard bound for the tail of a Gaussian distribution (for any $z \geq 0$ we have $1 - \Phi(z) = \Phi(-z) \leq \frac{1}{2}e^{-z^2/2}$). Putting all this together, we get that number of errors of the procedure can be upper bounded as

$$\begin{aligned}
\mathbb{E}_S[\hat{S}\Delta S] &= \sum_{i\in S}\mathbb{P}(i \notin \hat{S}) + \sum_{i \notin S}\mathbb{P}(i \in \hat{S}) \qquad (4) \\
&\leq \frac{sT}{2}\exp(-\Gamma\mu^2/2) + (n-s)\,2^{-T}.
\end{aligned}$$

Now all that we have left is to choose the parameters Γ and T such that equation (1) is satisfied and that the error is small, say smaller than a pre-determined level $\varepsilon > 0$. Note that this is a balancing act: on one hand we would like to choose the precision Γ and the maximal number of measurements T large, so that the error becomes small, as shown by inequality (4). However, doing so increases the amount of precision used by the procedure, as shown by inequality (3).

Note that the second term on the right hand side of inequality (4) only involves T. Hence, if we want the error to be at most ε than T has to be chosen to make that term strictly smaller than ε, say $\varepsilon/2$. This leads to the choice $T = \log_2(2(n-s)/\varepsilon)$. However, once we have T chosen, Γ is determined by the need to satisfy the precision budget. In particular, using the choice $\Gamma = \frac{m}{3n}$ with inequality (3) yields

$$\mathbb{E}\left(\sum_k \Gamma_k\right) \leq m\left(\frac{2(n-s)}{3n} + \frac{s\log_2(2(n-s)/\varepsilon)}{3n}\right).$$

Note that the first term in the brackets above is at most $\frac{2}{3}$. The second term in the brackets is a function of n, s and ε. However, we are considering scenarios where n is very large, and s is much smaller than n. In such cases, unless ε is chosen to be extremely small, this term is at most $\frac{1}{3}$. Stated differently, for a given $\varepsilon > 0$ and assuming $s \leq n/(\log_2(n))$ equation (1) is satisfied provided n is large enough.

The last piece of the puzzle is to figure out how large the signal strength μ needs to be so that the error of the procedure above is at most ε. We have seen that our choice of T ensures that we do not erroneously include zero components in $\hat{S}$, but unless the signal is strong enough we cannot expect to correctly identify non-zero components. Formally, we see from inequality (4) that to ensure an error of at most ε we need

$$\frac{sT}{2}\exp(-\Gamma\mu^2/2) \leq \frac{\varepsilon}{2}.$$

Rearranging the above, and plugging in values for T and Γ yields

$$\mu \geq \sqrt{\frac{6n}{m}\left(\log\frac{s}{\varepsilon} + \log\log_2\frac{2(n-s)}{\varepsilon}\right)}.$$

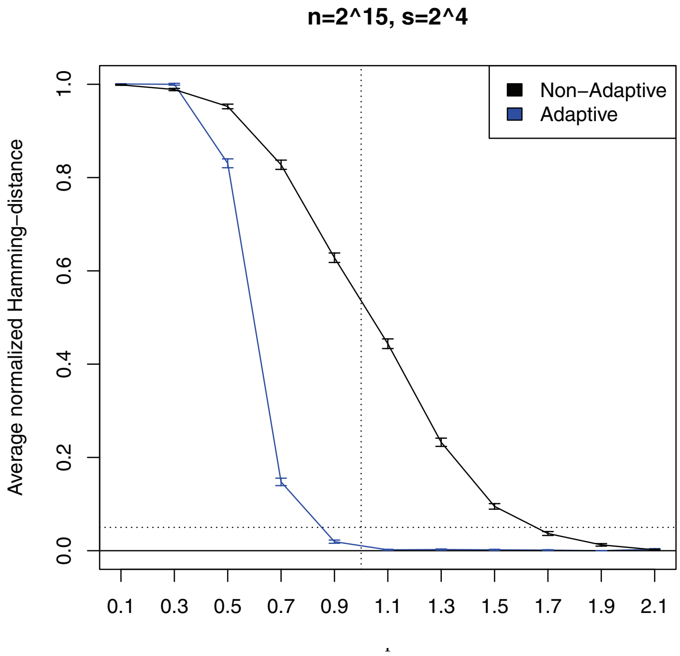

Figure 1 Average normalized error (with SE bands) for the different estimators as a function of the parameter t (the signal strength is $t \cdot \mu_{\text{limit}}$): the non-adaptive estimator(black); adaptive sensing estimator (blue). The number of repetitions is 100 for each value of t. The vertical black dashed line is at the value $t = 1$. The horizontal black dashed line is at the value of $\varepsilon = 0.05$.

In other words we have shown that whenever the signal strength satisfies the previous inequality, the procedure described above has error at most ε.

A few remarks are now in order. We first remark that the $\log\log n$ term in the bound above is an artifact of the simple method we presented, and one can use a more refined procedure to eliminate it from the bound. The same is true for the constant 6 in the bound, which can be improved to a value of 2. Actually, for each coordinate, the procedure described is essentially a poor-man's version of the celebrated Sequential Likelihood Ratio (SLR) test [24]. The latter is significantly more involved, but its performance is qualitatively the same.

To summarize, there is a slightly more sophisticated procedure which can achieve error no larger than ε provided that $\mu \geq \sqrt{\frac{2n}{m}\log\frac{s}{\varepsilon}}$. Contrasting this with the non-adaptive bounds presented before, we see that we can improve the $\log n$ factor to a $\log s$ factor in the bounds for μ using an adaptive procedure. In particular, this adaptive procedure is able to identify supports of signals that are so weak that doing so with any non-adaptive procedure is completely hopeless.

In Figure 1 we give a comparison of non-adaptive and adaptive sensing based on simulation. We plot the performance of the adaptive and non-adaptive procedures for different values of the signal strength μ. Theory tells us that the error of the adaptive procedure should drop below the value ε when $\mu = \mu_{\text{limit}} := \sqrt{\frac{2n}{m}\log\frac{s}{\varepsilon}}$. Hence we run the procedures for signal strength $\mu = t \cdot \mu_{\text{limit}}$ with values of t varying around 1. Our other parameter choices are $n = 2^{15}, m = n$ and $s = 2^4$. For each value of t we run both methods 100 times and plot their average error along with error bars whose total length is four times the (point-wise) standard error, which would correspond to a roughly 95% two-sided confidence interval for normally distributed measurements. Finally we note that instead of the simple procedure described in this manuscript, we use the somewhat more sophisticated test described in [8] that replaces the simple thresholding procedure with SLR tests, as described above. However, the results do not differ substantially when using the simple thresholding procedure we described. As can be seen in Figure 1 adaptive sensing clearly outperforms non-adaptive sensing, for the same precision budget, as expected.

Structured support estimation

We have seen that adaptive support estimation procedures have a marked advantage over non-adaptive procedures. In particular, when $m = n$ adaptive procedures can identify the support of sparse signals whose strength scales as $\sqrt{\log s}$, whereas the weakest sparse signal any non-adaptive procedure can deal with needs to scale as $\sqrt{\log n}$. The gap between the performance of adaptive and non-adaptive procedures can even be more dramatic, if the support of the signal is known to have some structural properties.

Assuming the signal support has some structure is a reasonable assumption in many applications. For instance, in gene expression studies one knows the expression levels of certain genes tend to be correlated, giving rise to interval-like structures in the gene-expression vector. In the same context, if one stacks the gene-expression levels of different individuals into a matrix, one expects to see that individuals with the same phenotype (i.e. having the same medical condition) have similar genes expressed, giving rise to sub-matrix structures in the gene-expression matrix. As another example, while monitoring infections on a network (be it a biological infection over a geographic region or the spreading of malware on a network of computers) we may expect to see star-shaped patterns of anomalous behavior radiating from the point of origin of the infection.

Structural assumptions are naturally incorporated in the model by restricting the class C to contain only sets with specific structural properties. For instance, when thinking of interval structures in signal vectors, we might restrict C to contain only sets that consist of s consecutive elements of the vector, instead of every possible set with cardinality s as done earlier. Similarly, if we would like to consider star-shaped activations of size s in a network, we would simply restrict C to contain sets of size s corresponding to star-shaped patterns. More precisely, suppose there is a graph $G = (V,E)$ encoding our network, and we want to consider star-shaped edge activations in G. Then we simply construct a map $\psi : E \rightarrow \{1,\ldots,n\}$ that identifies each edge of the network with a number between 1 and n. Then C consists of elements $\phi(S)$, where S is a star-shaped edge pattern in the network graph.

Under such assumptions, one might want to tailor their support estimation procedure to be more sensitive to such activation patterns, hopefully being able to identify even weaker signals. For instance we might imagine that we would scan the signal for activations of a specific structural form only. The hope is that since we are considering only a restricted set of activation patterns, we might be able to 'focus' our attention better, resulting in improved performance.

Non-adaptive sensing. We can naturally adapt non-adaptive procedures to cope with structural assumptions. We are primarily interested in situations where, a priori, no component of the signal is more likely to be active. This is formally stated as the assumption that the class C is symmetric, as in Definition 1. As discussed before, for symmetric classes there is no reason to measure any signal component with more precision than any other. In other words, the optimal non-adaptive sensing scheme still measures every component once, with the same precision m/n, and the non-adaptive estimator also remains unchanged and is given by (2). However, in some situations the set C under consideration is much smaller than before, which can benefit the performance of the estimator, as illustrated by Proposition 1.

Considering the class of intervals of size s, the size of the class is $|C| \approx \frac{n}{s}$ and the bound becomes $\sqrt{\log \frac{n}{s}/s}$. This is a significant improvement over the case of unstructured supports. This arises from the fact that the class of intervals of size s is much smaller than the class of all supports of size s. Unfortunately, the possibility for improvement vanishes if the class under consideration contains too many sets, even if the sets themselves have considerable structure to them. An example of this is the class of star-shaped activations in a network modeled by a complete graph. The number of star-shaped edge patterns in a complete graph $G = (V,E)$ is $|C| = |V| \binom{|V|-1}{s} \approx (\sqrt{n}/s)^s$, since we have $|V|$ choices for the center of the star and $\binom{|V|-1}{s}$ choices for the edges, given the center. This results in the bound of the order $\sqrt{\log \frac{n}{s}}$ which is the same as the bound for the unstructured case. In words, even though there is considerable structure to the sets in this class non-adaptive are not able to capitalize on that.

Adaptive sensing. In stark contrast with the above observations is the performance of adaptive sensing procedures. It turns out that the cardinality of the class C no longer plays the crucial role in the performance. A completely general characterization of the performance of adaptive procedures for structured classes is still missing, but we nevertheless have a general way to approach the problem.

A way to construct adaptive sensing procedures is to first consider the problem without observation noise. In such a case a reasonable approach is to first search across the signal vector until we find an element of the support. Once that happens, we can transition into an exploitation phase to find 'nearby' elements of S, taking advantage of any local structure the support might have. Once the local structure has been exploited, the algorithm can revert back to the exploration phase to find another signal component, if there is a part of S that has not been found yet.

Once we have a method for noiseless observations, we can make it robust to noise by repeating each 'noiseless observation' with hypothesis tests to determine the identity of the component in question. These tests can be very similar to the one described previously, or simply SLR tests. In order to get a reliable procedure, the tests need to be properly calibrated — a full description of the approach is too involved to be presented here, but it is fully outlined in [8].

As an example, consider a simple procedure for the case of interval activations. This procedure will consist of two phases. The first phase will be designed to find one component of the signal support S. We call this the exploration phase. Once a component of S has been found, we move on to the exploitation phase, in which we find the remaining elements of S by sampling in the vicinity of the previously found component.

It can be shown that, for the class of intervals with $s < \sqrt{n}$, such a procedure has error smaller than ε provided $\mu \geq \sqrt{\frac{2}{s} \log \frac{2s}{\varepsilon}}$. Conversely, recovering the support is impossible for weaker signals. So, for the class of intervals adaptive sensing has a gain over adaptive sensing by replacing the factor $\sqrt{\log n}$ by $\sqrt{\log s}$. This is relatively modest.

Now consider the class of star-shaped sets. In that case, one can show that an adaptive sensing procedure (very similar to the one above) has error smaller than ε provided $\mu \geq \sqrt{\frac{2}{s} \log \frac{4s}{\varepsilon}}$ (this result assumes s slightly smaller than $\sqrt{n}$). In contrast any non-adaptive procedure needs the signal magnitude to have order $\sqrt{\log \frac{n}{s}}$, which is dramatically higher and essentially the same performance had we not taken structure into consideration. In summary, adaptive sensing is able to capitalize on structural assumptions to a much greater extent than non-adaptive sensing.

In the next section we encounter a similar phenomenon, but in a different context — that of signals that change during the measurement period.

Dynamically evolving signals

In the previous sections we considered situations where the observed phenomenon did not change during the measurement process. However, in certain applications this assumption is not completely realistic. Consider for instance a signal intelligence setting where one wishes to detect covert communications. Suppose that our task is to survey a signal spectrum, a small fraction of which may be used for communication, meaning that some frequencies would exhibit increased power. On one hand we do not know beforehand which frequencies are used, but also the other parties may change the frequencies they communicate through over time. This means we will be chasing a moving target. This introduces a further hindrance in our ability to detect whether someone is using the surveyed signal spectrum for covert communications. Other motivating examples for such a problem include spectrum scanning in a cognitive radio system [4,18], detection of hot spots of a rapidly spreading disease [19,22,25,26], detection of momentary astronomical events [23] or intrusions into computer systems [14,21].

Signal model

To investigate such settings we first need to develop a model that captures the dynamical nature of such signals. As a start, our aim is to create a model that resembles the previous one as much as possible, so that we can clearly isolate the effect the dynamics has on the problem. Let $x^{(t)}$ denote the signal at time t, where $t = 1,2,\ldots$. At each time step, the signal is of the form

$$x_i = \begin{cases} \mu & \text{if } i \in S^{(t)}, \\ 0 & \text{if } i \notin S^{(t)}, \end{cases}$$

where $S^{(t)}$ is the support of the signal at time t. The support is always a set of

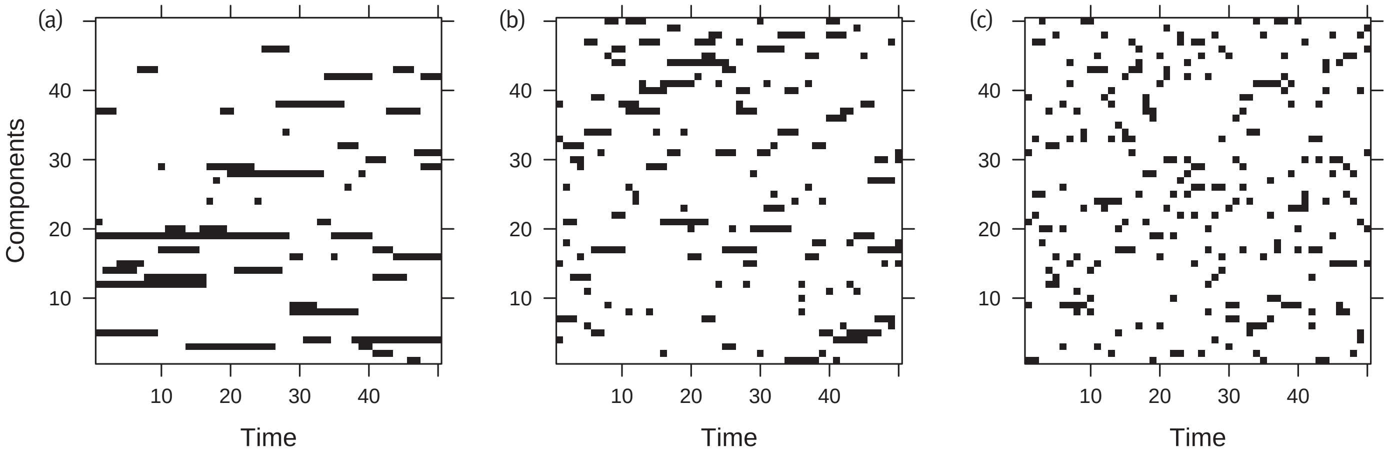

Figure 2 Simulation of the support dynamics with $n=50$, $s=5$ and (a) $p=0.2$; (b) $p=0.5$; and (c) $p=0.8$. Components in the support are colored black.

cardinality s as before, but the identity of the elements of the support changes over time. To initialize, $S^{(1)}$ is chosen uniformly at random among all sets of size s. For later times $t=2,3,\ldots$, in order to get the support $S^{(t)}$ from $S^{(t-1)}$, we flip a coin for each element of $S^{(t-1)}$ (a total of s flips). If the coin comes up heads, that element will also be included in $S^{(t)}$. On the other hand, if the coin comes up tails, that element 'moves' to a location chosen uniformly at random among the available locations (all those not corresponding to head-flips). The coins we flip comes up tails with probability p.

In this setup p dictates the speed of change of the signal. When $p=0$, the coin always comes up heads, and thus components never move, or in other words $S^{(t)}=S^{(t+1)}$ for all times t. In other words, the case $p=0$ is the same as the unstructured model of the previous section. On the other extreme, when $p=1$ the coin always comes up tails, therefore all components of the support 'move' at each time step. That is to say, when $p=1$ a new signal support is drawn uniformly at random at each time step. This dynamics are illustrated in Figure 2.

Measurement model and inference goal

As before, the signal cannot be observed directly, only through some sort of noisy measurement mechanism. We consider the same measurement model as before, but with the constraint that the precision of each measurement is 1. This is aligned with the view that the precision of a measurement is proportional to the time it takes to collect that measurement, as already mentioned in the section 'Measurement model'. Formally, the measurement model is

$$Y_t = x^{(t)}_{A_t} + W_t, \quad t=1,\ldots,m, \tag{5}$$

where A_t denotes the component of $x^{(t)}$ we measure at time t, and W_t is independent standard normal noise. Note that we are allowed to make a total of m measurements, which is a direct analogue of the precision budget constraint of equation (1).

Since the signal support might be changing between time steps, it is not clear what we would mean by support estimation in this setting. However, there is a closely related question one often asks in the context of sparse signals, which is well-defined even for dynamically evolving signals — this is the task of signal detection. Recall that support estimation equates to identifying components of the signal that exhibit anomalous activity. A statistically easier question is whether there is any anomalous activity at all? This can be formalized as a test between two hypotheses. Under the null, every component of the signal behaves nominally, or in other words $S^{(t)} \equiv \emptyset$ for every $t=1,\ldots,m$. Under the alternative, there is in fact anomalous activity in the signal, evolving according to the model described above. Our task is to decide which of these two hypotheses does our data support more.

Common sense tells us that unless we have enough time to monitor the system, there is no hope of reliably deciding between the two hypothesis, regardless of the signal strength μ. The reason is simple: since the support is sparse, there is very little chance at any given time of sampling an element of the support (under the alternative). To illustrate, imagine a situation where there is no measurement noise. In this case, the reasonable thing to do would be to collect samples of randomly selected components, until we find one whose value is non-zero. This would be hard evidence for the alternative hypothesis. If we sample for a long time, and never came across a non-zero entry, that would be evidence for the null. But how long is long enough? One can show that unless the number of measurements is of the order n/s, there is no hope for any procedure to reliably solve the signal detection problem, regardless of the signal strength μ or the speed of change p. Because of this, we call the setting when $m \approx n/s$ the *small sample regime* and this is the setup we focus on for now. If we are interested in a situation where our goal is to make a decision as fast as possible, this is the setting to consider.

Overview of results

We would like to understand how non-adaptive and adaptive sensing methods fare in the signal detection task described above. Recall that in non-adaptive sensing, the decision which components to measure needs to be made before the sampling process begins. It turns out that, in the small sample regime, static and fast-moving signals are equally hard to detect. One can formalize a necessary condition for the signal strength that needs to be satisfied for *any* non-adaptive method to work. Due to technical difficulties, a formal proof is so far only available in the extreme cases when $p=0$ and $p=1$. In both cases,

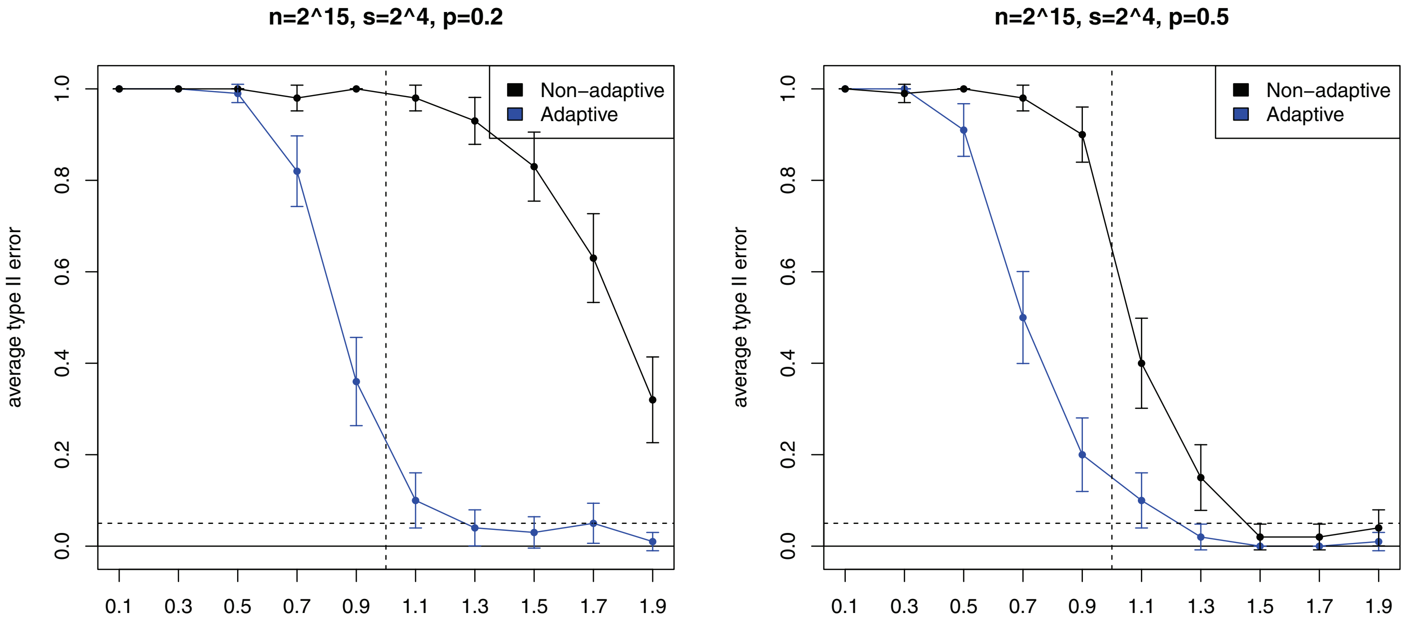

Figure 3 Average type II error probabilities (with SE bands) for the different estimators as a function of the parameter t (the signal strength is $t \cdot \mu_{\text{limit}}$ with $\mu_{\text{limit}} = \sqrt{4p\log(n/s)}$): the non-adaptive test (black); the adaptive sensing test based on the STT (blue). The plots from left to right correspond to $p = 0.2$ and $p = 0.5$. The number of repetitions is 100 for each value of t. The vertical black dashed line is at the value $t = 1$. The horizontal black dashed line is at the value of $\varepsilon = 0.05$.

the signal needs to be at least of order $\sqrt{\log\frac{n}{s}}$ for reliable detection to be possible by non-adaptive methods (see Theorem 4.1 of [10]). This shows that indeed static signals are as difficult to detect as constantly changing ones. Although a formal proof is missing, common sense leads one to believe that the detection boundary should follow the same scaling for $0 < p < 1$ as well.

We now turn our attention to adaptive sensing methods. As before, our hope is that the flexibility in the sampling process would translate to improved performance over non-adaptive methods. In particular, perhaps there is a way to apply a similar strategy to the one we saw before, i.e. to quickly discard zero components and focus sensing effort on non-zero components. Indeed this can be done, using a test with similarities to the one described earlier. Note, however, that we need also to take into account the signal is dynamic, which means that a non-zero component might move away while we are collecting samples of it. Therefore, we need to make our decisions fast, while also keeping our accuracy high. This can be achieved by subtle changes to the sequential test described earlier — we refer the interested reader to [10].

Once we have a test that can quickly and reliably identify whether a signal component is zero or not, the detection procedure is quite simple. Our strategy is to probe roughly $\frac{n}{s}$ components of the signal one after the other. When probing a component, we take repeated measurements of it, and apply our sequential test to decide if that component is zero or not. If the sequential test is designed well, it will be able to make an accurate and quick decision. If the test deems a component non-zero, we stop sampling and decide that the alternative hypothesis is true. However, if all probed components are deemed zero by the test, we decide that the null is true.

It can be shown that this procedure returns the correct decision with probability $\geq 1-\varepsilon$ whenever the signal strength μ scales as $\sqrt{\max\{p\log\frac{n}{s}, \log\frac{1}{\varepsilon}\}}$, where recall that p parametrizes the speed of change of the signal. That is to say, in the worst case when the signal is constantly changing (i.e. $p = 1$), the performance of adaptive and non-adaptive procedures are the same, and there is nothing to be gained. However, the adaptive procedure does take advantage of situations when the signal only changes slowly. Remarkably, in the extreme case of static signals ($p = 0$), the detection boundary no longer depends on n!

We illustrate these points by a simple numerical experiment, similar to the one presented in the previous section. As mentioned above, the error of the adaptive sensing procedure should drop below the value ε when the signal strength is around $\mu_{\text{limit}} = \sqrt{4p\log\frac{n}{s}}$ (the constant 4 comes from the detailed computations in [10]). Therefore we run both adaptive and non-adaptive tests in situations when the signal strength is $t \cdot \mu_{\text{limit}}$ with different values of t (around 1).

We run both procedures 100 times for every setting of t, and plot the average errors along with error bars in Figure 3. The error bars are constructed the same way as in the previous section. Our choice for the remaining problem parameters are $n = 2^{15}, s = 2^4$ as before, but now we consider the small sample regime $m = \frac{n}{s}\log_2(\frac{2}{\varepsilon})$ (the details behind the exact choice of m can be found in [10]). We run two different simulations, first when the speed of change p is equal to 0.2 and second when it is equal to 0.5.

As expected, the adaptive sensing test outperforms the non-adaptive one. The margin by which the adaptive sensing test is superior is greater as p decreases, which is what the theory suggests. Note that the error probability of the tests never quite reach the value zero, since with this choice of m, the probability of being able to sample an active component is not overwhelming.

Final remarks

In this paper we hope to have given the reader an idea of the potential of adaptive sensing, as well as the methodologies required for the development of sound algorithms. As mentioned earlier, the signal and measurement models can be significantly extended, going beyond Gaussian noise and purely sparse signals. Furthermore, instead of considering coordinate-wise measurements one can also consider linear ensembles — this is known as compressive sensing. Although technically more involved, similar results to the ones described here can be obtained in that setting as well [9] (although some open questions remain).

A related problem to the detection/estimation of means is the detection of correlations, where the aim is to determine whether there are correlated components within a high-dimensional vector [6]. In that scenario coordinate-wise measurements are not informative, and one always needs to sample multiple components simultaneously to determine if they are correlated. In that setting adaptive sensing can be advantageous when there is significant structure to the set of correlated components. However, it is an open question if in unstructured cases there is still a significant advantage to using adaptive sensing techniques.

References

1 Louigi Addario-Berry, Nicolas Broutin, Luc Devroye and Gabor Lugosi, On combinatorial testing problems, *The Annals of Statistics* 38(5) (2010), 3063–3092.

2 Maria-Florina Balcan, Alina Beygelzimer and John Langford, Agnostic active learning, *Journal of Computer and System Sciences* 75(1) (2009), 78–89.

3 S. Bubeck and N. Cesa-Bianchi, *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning, Now Publishers, 2012.

4 Raied Caromi, Yan Xin and Lifeng Lai, Fast multiband spectrum scanning for cognitive radio systems, *IEEE Transactions on Communications* 61(1) (2013), 63–75.

5 Rui M. Castro, Adaptive sensing performance lower bounds for sparse signal estimation and testing, *Bernoulli* 20(4) (2014), 2217–2246.

6 Rui M. Castro, Gabor Lugosi and Pierre-Andre Savalle, Detection of correlations with adaptive sensing, *IEEE Transactions on Information Theory* 60(12) (2014), 7913–7927.

7 Rui M. Castro and Robert D. Nowak, Minimax bounds for active learning, *IEEE Transactions on Information Theory* 54(5) (2008), 2339–2353.

8 Rui M. Castro and Ervin Tánczos, Adaptive sensing for estimation of structured sparse signals, *IEEE Transactions on Information Theory* 61(4) (2015), 2060–2080.

9 Rui M. Castro and Ervin Tánczos, Adaptive compressed sensing for estimation of structured sparse sets, *IEEE Transactions on Information Theory* 63(3) (2017), 1535–1554.

10 Rui M. Castro and Ervin Tánczos, Are there needles in a moving haystack? Adaptive sensing for detection of dynamically evolving signals, arXiv:1702.07899, to appear in *Bernoulli*, 2017.

11 David Donoho and Jiashun Jin, Higher criticism for detecting sparse heterogeneous mixtures, *Annals of Statistics* 32(3) (2004), 962–994.

12 Simon Foucart and Holger Rauhut, *A Mathematical Introduction to Compressive Sensing,* Birkhäuser, 2013.

13 Aurlien Garivier and Emilie Kaufmann, Optimal best arm identification with fixed confidence, in Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, eds., *29th Annual Conference on Learning Theory, 23–26 June 2016*, Proceedings of Machine Learning Research, Vol. 49, Columbia University, pp. 998–1027.

14 Robert Gwadera, Mikhail J. Atallah and Wojciech Szpankowski, Reliable detection of episodes in event sequences, *Knowledge and Information Systems* 7(4) (2005), 415–437.

15 S. Hanneke, *Theory of Disagreement-Based Active Learning*, Foundations and Trends(r) in Machine Learning Series, Now Publishers, 2014.

16 Jarvis Haupt, Rui M. Castro and Robert Nowak, Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory* 57(9) (2011), 6222–6235.

17 R. King, J. Rowland, S.G. Olivier, M. Young, W. Aubrey, E. Byrne, M. Liaka, M. Markham, P. Pir, L.N. Soldatova, A. Sparkes, K.E. Whelan and A. Clare. The automation of science, *Science* 324 (2009), 85–88.

18 Husheng Li, Restless watchdog: Selective quickest spectrum sensing in multichannel cognitive radio systems, *EURASIP Journal on Advances in Signal Processing* 2009(1), 417457.

19 Wuqiong Luo and Wee Peng Tay, Finding an infection source under the sis model, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 2930–2934.

20 Matthew L Malloy and Robert D Nowak, Sequential testing for sparse recovery, *IEEE Transactions on Information Theory* 60(12) (2014), 7862–7873.

21 Vir V. Phoha, *Internet Security Dictionary*, Springer Science & Business Media, 2007.

22 Devavrat Shah and Tauhid Zaman, Rumors in a network: who's the culprit? *IEEE Transactions on Information Theory* 57(8) (2011.), 5163–5181

23 David R. Thompson, Sarah Burke-Spolaor, Adam T Deller et al., Real-time adaptive event detection in astronomical data streams, *IEEE Intelligent Systems* 29(1) (2014), 48–55.

24 Abraham Wald. Sequential tests of statistical hypotheses, *The Annals of Mathematical Statistics* 16(2) (1945), 117–186.

25 Heng Wang, Minh Tang, Yu-Seop Park and Carey E. Priebe, Locality statistics for anomaly detection in time series of graphs, *IEEE Transactions on Signal Processing* 62(3) (2014), 703–717.

26 Kai Zhu and Lei Ying, Information source detection in the sir model: A sample path based approach, in *Information Theory and Applications Workshop* (ITA), 2013, pp. 1–9.