On Finding a Set of Healthy Individuals from a Large Population

Abhay Sharma and Chandra R. Murthy
Dept. of ECE, Indian Institute of Science, Bangalore 560 012, India
abhay.bits@gmail.com, cmurthy@ece.iisc.ernet.in

Abstract—In this paper, we explore fundamental limits on the number of tests required to identify a given number of “healthy” items from a large population containing a small number of “defective” items, in a nonadaptive group testing framework. Specifically, we derive mutual information-based upper bounds on the number of tests required to identify the required number of healthy items. Our results show that an impressive reduction in the number of tests is achievable compared to the conventional approach of using classical group testing to first identify the defective items and then pick the required number of healthy items from the complement set. For example, to identify \( L \) healthy items out of a population of \( N \) items containing \( K \) defective items, when the tests are reliable, our results show that \( O(K(L - 1)/(N - K)) \) measurements are sufficient. In contrast, the conventional approach requires \( O(K \log(N/K)) \) measurements. We derive our results in a general sparse signal setup, and hence, they are applicable to other sparse signal-based applications such as compressive sensing also.

I. INTRODUCTION

In typical sparse signal models, out of \( N \) input variables, only a small subset of variables of size \( K \ll N \) contribute to the observed output. For example, in a nonadaptive group testing setup [1], the group test outputs depend only on the columns of the test matrix corresponding to the defective items. Similarly, in a compressive sensing setup [2], the output signal is given by a linear combination of the columns of the measurement matrix corresponding to the nonzero entries of the input vector. This subset of salient inputs is referred to by different names, e.g., defective items, sick individuals, support set, etc. In the sequel, we will refer to it as the active set, and its complement as the inactive set. In the context of sparse signal recovery, active set identification is a problem that has received significant research attention in recent years. In contrast, this paper addresses the issue of the inactive subset recovery. To elaborate, we focus on the task of finding an \( L \leq N - K \)-sized subset of the inactive set (which is of size \( N - K \)), given the observations from a sparse signal model with \( N \) inputs, out of which \( K \ll N \) are active.

Finding a subset of items belonging to the inactive set is of interest in many applications. An example is the spectrum hole search problem in cognitive radio (CR) networks [3]. It is well known that the primary user occupancy is sparse in the frequency domain over a wide band of interest [4]. To setup a CR network, the secondary users need to find an appropriately wide unoccupied frequency band. Thus, the main interest here is the identification of only a subset of all the unoccupied bands, i.e., it is an inactive subset recovery problem. Furthermore, typically, the bandwidth of the required spectrum hole will only be a small fraction of the available free bandwidth. Another example is a product manufacturing setup, where a small shipment of non-defective items has to be delivered on a high priority. Once again, the interest here is on the identification of only a subset of the inactive items.

Motivated by the fact that the active set (e.g. primary occupancy) is sparse, a natural approach to finding a subset of inactive items is to first use efficient active set recovery methods, e.g., nonadaptive group testing, to first identify all the active items. This results in the identification of the full inactive set also, and one can then pick the required number of inactive items from this set. In this work, we ask whether this two-step approach is the best, or whether directly looking for an inactive subset would work better, in terms of the number of tests or measurements involved. This question is especially pertinent, when the size of inactive subset is small compared to the total number of inactive items.

Related work: In the context of sparse signal models, a large body of results now exist that characterize bounds on the number of measurements required for support recovery from noisy linear projections, both in information theoretic settings as well as for computationally tractable recovery methods (see [5]–[8] and the references therein). Similar results are available from the group testing literature that characterize the number of tests required to identify the defective items, both under information theoretic settings and for tractable decoding algorithms (see, e.g., [9], [10]). However, to the best of our knowledge, the question of finding an inactive subset given the observations of the sparse signal model has not been considered in the literature. The results presented in this paper are the first in this direction.

In this paper, we consider a general sparse signal model consisting of \( N \) input covariates, out of which an unknown subset of size \( K \) is active. Observations or outcomes are generated according to an observation model that only depends on the active covariates. This model was first proposed in [9], in the context of sparse signal support recovery. Given multiple observations from the above model, we propose decoding schemes to identify a subset of \( L \) inactive variables. We analyze two decoding schemes: (a) Identifying the active set and then choosing \( L \) inactive covariates randomly from the complement set, and (b) Directly identifying the inactive subset from the observations. By analyzing the average probability of error for the two decoding schemes, we derive the upper
bounds on the number of observations required to identify a set of \(L\) inactive variables. Our main contributions are as follows:

- We derive mutual information-based upper bounds on the number of observations required for identifying an \(L\)-sized subset of inactive variables with the average probability of error vanishing asymptotically in \(N\).
- We specialize the bounds for various noisy nonadaptive group testing scenarios and characterize the required number of group tests in terms of \(L, N\) and \(K\).

Our results show that, compared to the conventional approach of identifying the inactive subset from the complement of the active set, an enormous reduction in the number of observations can be obtained by directly looking for an \(L\)-sized inactive subset, especially when \(L\) is small compared to \(N - K\). Further, application to the nonadaptive group testing setup provides insight into the interplay between the observation and noise model and the required number of tests for successful identification of a subset of inactive items.

**Notation:** For any positive integer \(a\), \([a]\) denotes \([1, 2, \ldots, a]\). For any set \(A\), \(A^c\) denotes complement operation and \(|A|\) denotes the cardinality of the set. For any two sets \(A\) and \(B\), \(A \cap B = A \cap B^c\), i.e., elements of \(A\) that are not in \(B\). \(\{\}\) denotes the null set. Scalar random variables (RV) are represented by capital non-bold alphabets, e.g., \(\{Z_1, Z_3, Z_5, Z_6\}\) represent a set of 4 scalar RVs. If the index set is known, we use a compact notation to represent this by using the index set as sub-script. For example, the above set can be represented as \(Z_S\), where \(S = \{1, 3, 5, 8\}\). Bold-face letters represent vector random variables. For example \(Z_S\) denotes a set of 4 random vectors indexed by the set \(S\). Individual vector RVs are also denoted with an underline, e.g., \(\underline{z}\) represents a single random vector. \(B(q), q \in [0, 1]\) denotes the Bernoulli distribution with parameter \(q\). \(\set A\) denotes the indicator function and returns 1 if the event \(A\) is true, and returns 0 otherwise. \(x = O(y)\) implies \(x \leq By\) for some \(B > 0\).

The rest of the paper is organized as follows. Section II describes the signal model and problem setup. This is followed by the derivation of upper bounds on number of observations in Sec. III. The bounds are specialized to the case of nonadaptive group testing in Sec. IV. Concluding remarks are offered in Sec. V. Due to lack of space, we omit the proofs of the theorems in this paper; these will be provided in [11].

## II. Problem Setup

In this section, we describe the signal model and problem setup. Let \(X_{[N]} = X_1, X_2, \ldots, X_N\) denote a set of \(N\) independent and identically distributed input random variables (or items). Let each \(X_j\) belong to an alphabet denoted by \(\mathcal{X}\) and be distributed as \(Pr\{X_j = x\} = Q(x), x \in \mathcal{X}, j = 1, 2, \ldots, N\). For a group of input variables, e.g., \(X_{[N]}\), \(Q(X_{[N]}) = \prod_{j \in [N]}^{} Q(X_j)\) denotes the joint distribution for all the input variables. We consider a sparse signal model where only a subset of the input variables are active (or defective), in the sense that only a subset of the input variables contribute to the output. Let \(S \subset [N]\) denote the set of input variables that are active, with \(|S| = K\). Let the output signal \(Y\) belong to an alphabet denoted by \(\mathcal{Y}\). We assume that \(Y\) and is generated according to a known conditional distribution \(P(Y|X_{[N]}\). Then, in our observation model, we assume that given the active set, \(S\), the output signal, \(Y\), is independent of the other input variables. That is, \(P(Y|X_{[N]} = P(Y|X_S) \forall Y \in \mathcal{Y}\). We observe the outputs corresponding to \(M\) independent realizations of the input variables and denote the inputs and the corresponding observations by \(\{X, Y\}\). Here, \(X\) is an \(M \times N\) matrix, with its \(i\)th row representing the \(i\)th realization of the input variables and \(Y\) is an \(M \times 1\) vector with its \(i\)th component representing the \(i\)th observed output. Let \(S^c = [N] \setminus S\) denote the set of variables that are inactive. Let \(L \leq N - K\). Given the observation set \(\{X, Y\}\), we wish to find a set \(S_H \subset S^c\) such that \(|S_H| = L\). In particular, our goal is to obtain information theoretic bounds on the number of observations, \(M\), required to identify a set of \(L\) inactive variables with average probability of error going to zero asymptotically in \(N\). Here, an error event is said to occur if the chosen inactive set contains one or more active variables.

The above signal model, first proposed in [9], is a generalization of some of the popular nonadaptive measurement models, e.g., compressed sensing, nonadaptive group testing etc. For example, by choosing \(\mathcal{X} = \{0, 1\}\), \(\mathcal{Y} = \{0, 1\}\) and choosing \(\mathcal{y}\) to be a logical-OR of all the columns of \(X\), we get the classical nonadaptive group testing setup with no noise. Similarly, different types of noise models can be incorporated, e.g., additive noise, dilution etc., by choosing an appropriate model for \(P(Y|X_S)\) [9]. Thus, using this model allows us to derive general mutual information based bounds on number of observations to find a set of inactive items, that are applicable in a wide variety of scenarios.

Some further comments about the signal model:

- The independence assumption across the input variables and across different measurements imply that each entry in \(X\) is independent and identically distributed.
- Let \(S\) be the given defective set. For any \(1 \leq j \leq K\), let \(S(j)\) and \(S(K-j)\) represent a partition of \(S\) such that \(S(j) \cup S(K-j) = S\), \(S(j) \cap S(K-j) = \{\}\) and \(|S(j)| = j\). Define \(I(j) = I(Y, X_{S(K-j)}; X_{S(j)})\) as the mutual information between \(Y, X_{S(K-j)}\) and \(X_{S(j)}\) [12]. Mathematically,

\[
I(j) = \sum_{Y \in \mathcal{Y}} \sum_{X_{S(K-j)} \in \mathcal{X}^{K-j}} \sum_{X_{S(j)} \in \mathcal{X}^j} P(Y, X_{S(K-j)}|X_{S(j)}) Q(X_{S(j)}) \log \frac{P(Y, X_{S(K-j)}|X_{S(j)})}{P(Y, X_{S(K-j)})} \tag{1}
\]

Using the independence assumptions in our signal model, we can see that for a given \(j\), this quantity is independent of the chosen defective set, \(S\), and of the partitions of \(S\). Hence, it is justifiable to denote it by \(I(j)\). When \(K = 1\), we have \(I(1) = I(Y; X_l)\) for any \(l \in [N]\).

In the next section, we derive upper bounds on the number of observations required to identify \(L\) inactive variables in the above signal model.
III. SUFFICIENT NUMBER OF OBSERVATIONS

We now present bounds on the number of observations that are sufficient to find an \( L \)-sized set \( S_H \) of inactive variables. Our general methodology is as follows: (a) Given a set of inputs and observations, \( \{ \mathbf{X}, \mathbf{y} \} \), we first propose a decoding algorithm to find \( S_H \); (b) For the given decoding algorithm, we find (or upper bound) the average probability of error, where the probability is evaluated over the random set \( \{ \mathbf{X}, \mathbf{y} \} \). An error event occurs when the decoded set \( S_H \) contains one or more defective variables, i.e., when \( S \cap S_H \neq \emptyset \); (c) We find the relationships amongst \( M, N, L \) and \( K \) that will asymptotically drive the average probability to zero. For the asymptotic results, we have assumed that \( K \) remains fixed while \( N \to \infty \). Section III-A describes the straightforward decoding scheme where we find the inactive variables by first isolating the active set followed by choosing the inactive set randomly from the complement set. This is followed by the analysis of a new decoding scheme we propose in Section III-B, where we directly search for an inactive subset of the required cardinality.

A. Decoding scheme 1: Look into the complement set

One straightforward decoding scheme to find a set of inactive variables is to first decode the defective set and then pick a set of \( L \) variables uniformly at random from the complement set. Here, we employ Maximum Likelihood (ML) decoding [9] to find the defective set. The probability of error in identifying the defective set with ML decoding was analyzed in [9]. Intuitively, even if the ML decoder incorrectly identifies the defective set, there is still a “high” probability of picking a correct inactive set, since in the complement set there remain only a few active variables. Hence, the error probability analysis of using this scheme to identify a non-defective subset is similar, with an extra term to account for the probability of picking an incorrect set of \( L \) variables from the complement set. Such an event happens with nonzero probability whenever the ML decoder picks an incorrect defective set. For this decoding scheme, we present the following result as a corollary to (Theorem III.1, [9]).

**Corollary 1.** Let \( N, M, L \) and \( K \) be as defined above. If

\[
M > \max_{1 \leq j \leq K} \log \left( \frac{(N-K)}{(N-L)} \binom{L}{j} \right) \cdot \frac{C_0(L, N, K, j)}{I(1)},
\]

where \( C_0(L, N, K, j) \) is as defined above, and

\[
I(1) = \sum_{i=1}^{K} \binom{N-K-j}{i},
\]

then, for all fixed \( K \geq 1 \), the average probability of error in finding \( L \) inactive variables, \( P_e \), approaches zero asymptotically in \( N \), i.e., \( \lim_{N \to \infty} P_e = 0 \).

Thus, we have obtained a bound on the number of observations that are sufficient to find a set of \( L \) inactive variables. Since \( C_0 \leq 1 \), the bound in (2) is better than the bound of using the same number of observations as is required to find the defective set [9]. Can we do better? The key idea, as we discuss in the next section, is to look at the problem independently of the problem of finding the full defective set.

B. Decoding Scheme 2: Find the inactive subset directly

For simplicity of exposition, we describe this decoding scheme in two stages. First, we present the result for the \( K = 1 \) case, i.e., when there is only one defective item. We later generalize the result to \( K > 1 \). The \( K = 1 \) case brings out the fundamental difference between finding defective items and finding non-defective items.

1) The \( K = 1 \) Case

We propose the following decoding scheme:

- Given \( \{ \mathbf{X}, \mathbf{y} \} \), compute \( P(\mathbf{y} | \mathbf{x}_i) \) for all \( i \in [N] \) and sort them in descending order. Since \( K = 1 \), we know \( P(\mathbf{y} | \mathbf{x}_i) \) for all \( i \in [N] \), and hence \( P(\mathbf{y} | \mathbf{x}_i) \) can be computed using the independence assumption across different observations.
- Pick the last \( L \) indices in the sorted array as the set of \( L \) inactive variables.

Note that, in contrast to finding defective set, the problem of finding \( L \) inactive variables does not have unique solution (except when \( L = N - K \)). Although the proposed decoding scheme provides a way to pick a solution, the probability of error analysis should take into account the fact that an error event happens only when the inactive set chosen by the decoding algorithm contains an active variable. The result is presented as the following theorem.

**Theorem 2.** Let \( N, M, L \) and \( K \) be as defined above, and let \( K = 1 \). If

\[
M > \frac{\log \left( \frac{N-K}{N-L} \right)}{(N-L)I(1)},
\]

where \( I(1) \) is as defined in (1), the average probability of error in finding \( L \) inactive variables, \( P_e \), satisfies \( \lim_{N \to \infty} P_e = 0 \).

We now generalize our decoding scheme to \( K \geq 1 \).

2) The \( K \geq 1 \) Case

It is easy to see that the decoding scheme for \( K = 1 \) does not directly extend to the \( K > 1 \) case. For example, let us arrange \( P(\mathbf{y} | \mathbf{x}_{S_d}) \) in decreasing order for all \( S_d \subset [N] \) such that \( |S_d| = K \). Since different \( S_d \) are not necessarily disjoint, it is possible that the sets towards the end of the sorted list would have overlapping entries. Thus, it is not clear how many entries from the rear we need to choose, in order to collect \( L \) distinct inactive variables. Hence, we propose to use a multi-stage algorithm, described below. At each stage, we find \( K \) variables that are most likely to be inactive, and remove them from consideration in the next stage. Later, from the probability of error analysis, we argue that such a decoding scheme has an asymptotic optimality property (see remarks (a) and (b) below).

**Decoding Scheme:**

- Initialize \( T_1 = [N] \); \( S_H = [ ] \).
- For \( i = 1, 2, \ldots, \left[ \frac{M}{L} \right] \) do:
  - Given \( \{ \mathbf{X}, \mathbf{y} \} \), compute \( P(\mathbf{y} | \mathbf{x}_{S_d}) \) for all \( S_d \subset T_i \) and \( |S_d| = K \). Find:
\[
S^{(i)}_{\omega} = \arg\min_{S_\omega \subset T_i \cap |S_\omega| = K} P(Y | X_{S_\omega}) \tag{4}
\]

- Set \(S_H = [S_H S^{(i)}_{\omega}]\) and \(T_{i+1} = T_i \setminus S^{(i)}_{\omega}\).

The probability of error analysis for the above multi-stage algorithm leads to a sufficient condition on the number of observations required for finding \(L\) inactive variables. We summarize the result in the following theorem.

**Theorem 3.** Let \(N, M, L\) and \(K\) be as defined above. If

\[
M > \frac{\log \left( \frac{(N-K)}{L_1 I^{(1)}} \right)}{L_2 I^{(1)}},
\]

where \(L_1 = N - \left( \left\lfloor \frac{L}{K} \right\rfloor + 1 \right) K + 1, \tag{5}\)

then, for all fixed \(K \geq 1\), the average probability of error in finding \(L\) inactive variables, \(P_e\), satisfies \(\lim_{N \to \infty} P_e = 0\).

We make the following comments about the above result:

(a) For \(K = 1\), (5) reduces to the bound in Theorem 2.

(b) We can use the above decoding scheme to find the active set also, by setting \(L = N - K\), i.e., by finding all the inactive variables. Suppose \(N - K\) is an integer multiple of \(K\). In this case, it is interesting to note that the above bound matches with the bound for finding \(K\) defective items that is based on the ML decoding rule (see Theorem III.1, [9]). This shows that the above decoding scheme is asymptotically equivalent to the ML decoding rule for finding the defective set when \(N - K\) is an integer multiple of \(K\), at least in terms of the upper bound on the number of observations required for support recovery.

(c) Define, \(\Gamma_{ud} \triangleq \log \left( \frac{(N-K)}{L_1 I^{(1)}} \right)\), \(\Gamma_{u1} \triangleq \Gamma_{ud} + \log \left( \frac{\sum_{i=1}^{I^{(1)}} (N-K)^{(i)}}{L_2 I^{(1)}} \right)\) and \(\Gamma_{u2} \triangleq \log \left( \frac{(N-K)}{L_1 I^{(1)}} \right)\). The bounds in (2), (5) and in Theorem III.1 [9] are all of the form \(\Gamma_{ud}/I^{(1)}\), with \(\Gamma_{u1}, \Gamma_{u2}\) and \(\Gamma_{ud}\) being a shorthand notation for \(\Gamma_{ud}(N, L, K)\) for each of the three schemes, respectively. Expressing the bounds in this form allows us to compare the bounds by focusing only the multiplicative factor \(\Gamma_{ud}\) for each bound, as the application-dependent mutual information term \(I^{(1)}\) is common to all the bounds.

Figure 1 presents a comparison between the multiplicative factors. It is clear that the approach of directly finding an \(L\)-sized inactive set requires far fewer number of observations compared to the other approaches.

(d) Consider \(\Gamma_{u2}\), i.e., the multiplicative factor in the upper bound for the decoding scheme 2 (see (5)). Define \(\alpha = \frac{\Gamma_{u2}}{\Gamma_{ud}} \in [0, 1]\). It can be shown that \(\Gamma_{u2}\) depends on \(N, L\) and \(K\) only as a function of \(\alpha\) (See Fig. 2). In particular, the size of inactive set to be found, \(L\), impacts the sufficient number of observations only through \(\alpha\). This is intuitively satisfying, since \(\alpha\) is the fraction of inactive variables that need to be found.

(e) From Fig. 1, we see that for low to moderate \(L\), the number of observations grows linearly and “slowly” with \(L\) (or \(\alpha\)). It is for these values of \(L\) that the direct approach provides significant reduction in the number of observations compared to the conventional scheme.

(f) In addition to the above upper bounds, it is possible to employ a modified version of Fano’s inequality to derive lower bounds, or necessary conditions, on the number of measurements required for identifying an \(L\)-sized subset of inactive items. We present these results in the extended version of this work [11].

In the next section, we characterize the exact scaling laws governing \(M, L, N\) and \(K\) in the context of a nonadaptive group testing application.

**IV. Finding Healthy Individuals Via Group Testing**

In this section, we apply the above mutual information based results to the specific case of nonadaptive group testing and characterize the number of tests that are sufficient to
identify a subset of healthy items from a large population. In a group testing framework [1], [9], we have a population of \(N\) items, out of which \(K\) are defective. Let \(G \subseteq [N]\) denote the defective set, such that \(|G| = K\). The group tests are defined by a boolean matrix, \(X \in \{0, 1\}^{M \times N}\), that assigns different items to different group tests (pools). In the \(i\)th test, the items corresponding to the columns with 1 in the \(i\)th row of \(X\) are tested, for \(i = 1, 2, \ldots, M\). We consider a random Bernoulli measurement matrix, where each \(X_{ij} \sim \mathcal{B}(p)\) for some \(0 < p < 1\), as in [9]. Here, \(p\) is a design parameter that controls the average group size. If the tests are completely reliable, then the output of the \(M\) tests is given by the boolean OR of the columns of \(X\) corresponding to the defective set \(G\).

In group testing, two different noise models are considered [9], [10]: (a) An additive noise model, where there is a nonzero probability \(q > 0\) that the outcome of a group test containing only non-defective items comes out positive; (b) A dilution model, where there is a nonzero probability \(u > 0\) that a given defective item does not participate in a given group test. Let \(d_i \in \{0, 1\}^M\). Let \(d_i(j) \sim \mathcal{B}(1-u)\) be chosen independently for all \(j = 1, 2, \ldots, M\) and for all \(i = 1, 2, \ldots, N\). Let \(D_i \triangleq \text{diag}(d_i)\). Let \(\mathcal{G}\) denote the boolean OR operation. The output vector \(y \in \{0, 1\}^M\) can be represented as:

\[
y = \bigvee_{i=1}^{N} D_i x_i \bigwedge_{i \in \mathcal{G}} w_i
\]

where \(x_i \in \{0, 1\}^M\) is the \(i\)th column of \(X\), \(w \in \{0, 1\}^M\) is the additive noise, with the \(i\)th component \(w(i) \sim \mathcal{B}(q)\). Note that, for the noiseless case, \(u = 0, q = 0\). In an additive model, \(u > 0\) and \(q > 0\). In a dilution model, \(u > 0\), \(q = 0\). Our goal here is to find the number of tests required to identify an \(L\) sized subset belonging to \([N] \setminus G\) using \(y\), with vanishing probability of error as \(N \to \infty\).

To compute the number of observations that are sufficient for finding an \(L\)-sized inactive subset, we use the lower bound on \(I^{(1)}\) derived in [9]. With \(p = \frac{1}{K}\) and by upper bounding the combinatorial term, we summarize the order- accurate sufficient number of tests to find a set of \(L\) healthy items in Table I. Here, we have considered \(L\) to be an integer multiple of \(K\) and that \(\alpha \triangleq \frac{L - 1}{N - K} \leq 0.5\). It is clear that our approach of directly identifying healthy individuals offers a significant advantage over the conventional approach of picking healthy items from the complement of the defective set. For example, with the direct approach, \(O\left(\frac{K(L-1)}{N-K}\right)\) tests are sufficient, while the indirect approach will require \(O(K \log \frac{N}{K})\) tests.

Note that, the results in Table I have been derived by setting \(\alpha = \frac{1}{K}\). These can be further improved by optimizing over the parameter \(p\), i.e., by using an optimal average group size. It is interesting to note that, although \(I^{(1)}\) depends on \(p\), the value of \(p\) that minimizes the number of tests does not depend upon \(L\). This is easily seen from the expressions for the sufficient number of tests, since \(L\) only impacts the combinatorial term, \(\Gamma_{uv}^2\), and not \(I^{(1)}\). Thus, the value of \(p\) can be chosen independently of \(L\).

\begin{table}[h]
\centering
\caption{Finding a subset of \(L\) healthy items: Order results for sufficient number of group tests when \(L \leq 0.5\).}
\begin{tabular}{|c|c|}
\hline
Noise Model & \(O\left(\frac{K(L-1)}{N-K}\right)\) \\
\hline
Additive noise & \(O\left(\frac{K(L-1)}{1-\alpha(1-L/N)}\right)\) \\
Dilution noise & \(O\left(\frac{K(L-1)}{1-\alpha(1-K/N)}\right)\) \\
\hline
\end{tabular}
\end{table}

\section{Conclusions}

In this paper, we considered the problem of identifying \(L\) healthy items out of a large population of \(N\) items containing \(K\) defective items in a general sparse signal modeling setup. We contrasted two approaches: identifying the defective items using the observations followed by picking \(L\) items from the complement set, and a directly identifying healthy items from the observations. We derived upper bounds on the number of observations required for identifying the \(L\) healthy items. We also applied the results in a nonadaptive group testing setup. We characterized the number of tests that are sufficient to identify a subset of healthy items of a large population under both dilution and additive noise models. We showed that an impressive gain in the number of observations is obtainable by directly identifying the healthy items. Our results were information theoretic in nature, without considering the practicality of the decoding algorithms. Future work could look at finding computationally tractable algorithms for directly identifying a subset of inactive variables.

\begin{thebibliography}{10}
\bibitem{1} D. Du and F. Hwang, \textit{Combinatorial group testing and its applications}, World Scientific, 1993.
\bibitem{2} E. J. Candès and T. Tao, “Decoding by linear programming,” \textit{IEEE Trans. Inf. Theory}, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
\bibitem{3} S. Haykin, “Cognitive radio: brain-empowered wireless communications,” \textit{IEEE J. Sel. Areas Commun.}, vol. 23, no. 2, pp. 201–220, Feb. 2005.
\bibitem{4} D. Cabric, S. M. Mishra, D. Willkomm, R. Brodersen, and A. Wolisz, “A cognitive radio approach for usage of virtual unlicensed spectrum,” in \textit{Proc. of 14th IST Mobile Wireless Communications Summit}, 2005.
\bibitem{5} S. Aeron, V. Saligrama, and M. Zhao, “Information theoretic bounds for compressed sensing,” \textit{IEEE Trans. Inf. Theory}, vol. 56, no. 10, pp. 5111–5130, Oct. 2010.
\bibitem{6} M. J. Wainwright, “Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting,” \textit{IEEE Trans. Info. Theory}, vol. 55, no. 12, pp. 5728–5741, 2009.
\bibitem{7} E. J. Candès, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” \textit{Communications on Pure and Applied Mathematics}, vol. 59, no. 8, pp. 1207–1223, 2006.
\bibitem{8} M. J. Wainwright, “Sharp thresholds for high-dimensional and noisy sparsity recovery using \(\ell_1\)-constrained quadratic programming (lasso),” \textit{IEEE Trans. Inf. Theory}, vol. 55, no. 5, pp. 2183–2202, May 2009.
\bibitem{9} G. Atia and V. Saligrama, “Boolean compressed sensing and noisy group testing,” \textit{IEEE Trans. Inf. Theory}, vol. 58, no. 3, pp. 1880–1901, 2012.
\bibitem{10} M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, “Group testing with probabilistic tests: Theory, design, and application,” \textit{IEEE Trans. Inf. Theory}, vol. 57, no. 10, pp. 7057–7067, Oct. 2011.
\bibitem{11} A. Sharma and C. R. Murthy, “On identifying a subset of healthy individuals in a large population,” In preparation, 2013.
\bibitem{12} R. G. Gallager, \textit{Information Theory and Reliable Communication}, John Wiley & Sons, Inc., New York, NY, USA, 1968.
\end{thebibliography}