On Finding a Subset of Healthy Individuals from a Large Population

Chandra R. Murthy
Joint work with Abhay Sharma
abhay.bits@gmail.com; cmurthy@ece.iisc.ernet.in

October 4, 2013

Sparse Signal Models

- Only a small subset of inputs contribute towards output
- Examples
  - Non-adaptive group testing
  - Compressive sensing
- Given the observed signal, recovery of input signals is well studied
  - Signal recovery
  - Support recovery
- Basic goal is to characterize the number of output observations required for recovery
  - Information theoretic limits
  - Performance for a computationally tractable method

“Healthy” Vs. “Sick” Individuals

- Typical non-adaptive group testing scenario
  - N individuals, $K \ll N$ are sick, $N - K$ are healthy
  - Multiple individuals are pooled in a single test
  - Goal: to identify all the sick individuals using as few group tests as possible
- For many applications, identification of a *healthy subset* is of prime importance
  - Identification of sick individuals is a straightforward but indirect way to find a subset of healthy individuals

"Healthy" Subset Identification: Examples

- Spectrum hole search in a cognitive radio network
  - Primary occupancy is sparse
  - Secondary users need to find only a "small free chunk"
    - A healthy subset identification problem!
    - Does the secondary network need to identify all the bands with primary occupancy?
- Entertaining a pushy customer!
  - Items manufactured with a small set of defectives
  - Need to urgently ship "a batch of non-defective items"
    - Do we need to identify all the defective items?
    - What is the minimum number of group tests required?
- Focus on identification of a subset of healthy items of a given size

Signal Model

- A set of N i.i.d. input RVs $(X_1, X_2, \ldots X_N)$
- An output Y generated according to a conditional distribution $P(Y|X_{[N]})$
- We consider the sparse signal model:
  - S_ω is the active (defective/sick) set, $|S_\omega| = K$
  - $P(Y|X_{[N]}) = P(Y|X_{S_\omega}), S_\omega \subset [N]$
    - Given the defective set, the output is independent of the other input variables
- We observe M outputs (denoted $\underline{\mathbf{y}}$), corresponding to M independent realizations of $X_{[N]}$ (denoted $\mathbf{X}$, size $M \times N$)

Problem Statement

- Given $\{\underline{\mathbf{y}}, \mathbf{X}\}$, find a set $S_\alpha \subset [N]$, such that

$$|S_\alpha| = L \text{ and } S_\alpha \cap S_\omega = \{0\}$$

  - Non-unique solutions
  - Recovery error if $S_\alpha \cap S_\omega \neq \{0\}$
- Goal: derive information theoretic limits for the number of observations, M, required to find L inactive variables

Summary of Approach

- Propose a decoding scheme to find L inactive variables
- Analyze the probability of error for the decoding scheme
- Find conditions on M, N, K, L such that the probability of error is exponentially decreasing in M
  - Scaling regime: $K \ll N$

Some Definitions: the E_0 Function

- 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 $|S^{(j)}| = j$
- For some positive integer n and $\rho \in [0, 1]$, define

$$E_0(\rho, j, n) = -\log \sum_{Y \in \mathcal{Y}} \sum_{X_{S^{(K-j)}} \in \mathcal{X}^{K-j}} \left\{ \sum_{X_{S^{(j)}} \in \mathcal{X}^j} Q(X_{S^{(j)}}) \left(P(Y, X_{S^{(K-j)}} | X_{S^{(j)}}) \right)^{\frac{1}{1+\rho n}} \right\}^{1+\rho n}$$

- The use of E_0 was pioneered by Gallager in characterizing the error exponents

Some Definitions: Mutual Information

- Define $I^{(j)} \triangleq I(Y, X_{S(K-j)}; X_{S(j)})$ as the mutual information between $\{Y, X_{S(K-j)}\}$ and $X_{S(j)}$

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

- Note that $\left.\frac{\mathrm{d}E_0(\rho, j, n)}{\mathrm{d}\rho}\right|_{\rho=0} = nI^{(j)}$

Approach 1: Look Into the Complement

- Finding all defective items solves our problem too!
- Decoding Scheme
  - Use an ML detection rule to find K defective items
  - Pick L items uniformly at random from the complement set

Corollary (To Thm. III.1 in [AtiaSaligrama12])

Let $C_0(L,N,K,j) \triangleq \frac{\sum_{i=1}^{j}\binom{N-K-j}{L-i}\binom{j}{i}}{\binom{N-K}{L}}$*. For any fixed* $K \geq 1$*, if*

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

then the average P_e *in finding L inactive variables* $\to 0$ *exponentially with the number of observations* M*.*

Approach 2, Take 1: Find Directly, $K = 1$

- Decoding scheme
  - Given $\{\mathbf{X}, \underline{\mathbf{y}}\}$, compute $P(\underline{\mathbf{y}}|\underline{\mathbf{x}}_i)$ for all $i \in [N]$
  - Arrange $P(\underline{\mathbf{y}}|\underline{\mathbf{x}}_i)$ in descending order
  - Pick the last L indices

Find Directly, $K = 1$: Probability of error

Theorem

Let $\rho \in [0\ 1]$.

$$P_e \leq \exp\left[-M\left(E_0(\rho, 1, N-L) - \frac{\rho \log \binom{N-1}{L-1}}{M}\right)\right].$$

Further, if

$$M > \frac{\log \binom{N-K}{L-1}}{(N-L)I^{(1)}},$$

then the average prob. of error in finding L inactive variables approaches zero exponentially with the number of observations.

Approach 2, Take 2: Find Directly, $K > 1$

- Decoding scheme for $K = 1$ does not extend directly
- A multi-stage decoding algorithm
  - Initialize $T_1 = [N]$; $S_H = [\,]$;
  - For $i = 1, 2, \ldots, \lceil \frac{L}{K} \rceil$ do:
    - Given $\{\mathbf{X}, \underline{\mathbf{y}}\}$, compute $P(\underline{\mathbf{y}}|\mathbf{X}_{S_\omega})$ for all $S_\omega \subset T_i$ and $|S_\omega| = K$. Find:

$$S_\omega^{(i)} = \underset{S_\omega \subset T_i, |S_\omega| = K}{\operatorname{argmin}} P(\underline{\mathbf{y}}|\mathbf{X}_{S_\omega})$$

    - Set $S_H = [S_H \; S_\omega^{(i)}]$ and $T_{i+1} = T_i \backslash S_\omega^{(i)}$
- Let $N_{\text{stg}} \triangleq \left\lceil \frac{L}{K} \right\rceil$, $L_j \triangleq (N - K) - (N_{\text{stg}} K - j)$

Find Directly, $K \geq 1$, Probability of error

Theorem

Let $C_2(L,N,K,j) \triangleq \binom{N-K}{L_j}\binom{KN_{stg}-j}{K-j}\binom{K}{1}\binom{K-1}{j-1}$. *Let* $\rho \in [0\ 1]$.

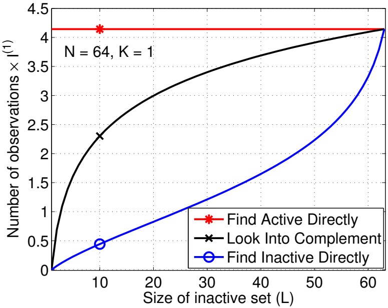
$$P_e \leq \sum_{j=1}^{K} \exp\left[-M\left(E_0(\rho,1,L_j) - \frac{\rho \log C_2(L,N,K,j)}{M}\right)\right].$$

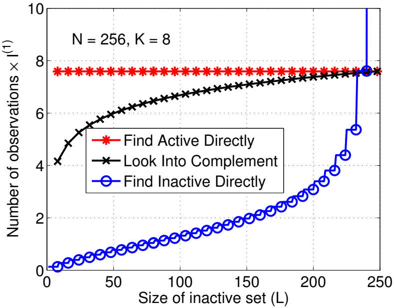
Further, if

$$M > \max_{1\leq j\leq K} \frac{\log C_2(L,N,K,j)}{L_j I^{(1)}},$$

then the average prob. of error in finding L inactive variables approaches zero exponentially with the number of observations.

Comparison of the Sufficient Number of Observations





- $\Gamma_{u2} \triangleq \max_{1\leq j\leq K} \frac{\log C_2(L,N,K,j)}{L_j}$, $\Gamma_{ud} \triangleq \log\left[\binom{N-K}{1}\binom{K}{1}\right]$ and $\Gamma_{u1} \triangleq \Gamma_{ud} + \log\left[\frac{\binom{N-K-1}{L-1}}{\binom{N-K}{L}}\right]$

Large N Behavior

- Linear in L for low to moderate values of L
- Asymptotic behavior
  - Let $\alpha \triangleq \frac{L-1}{N-K}$, fraction of the healthy items required
  - $\alpha \to 0$ as $N \to \infty$: L sub-linear in N
    - $\Gamma_{u2} \to 0$, $\Gamma_{u1} \to O(\log L)$, $\Gamma_{ud} \to O(\log N)$
  - $\alpha \to \alpha_0$ as $N \to \infty$: L linear in N
    - $\Gamma_{u2} \to \frac{H(\alpha_0)}{1-\alpha_0}$ (constant), $\Gamma_{u1}, \Gamma_{ud} \to O(\log N)$

Necessary Number of Observations

Theorem

Let N, M, L and K be as defined before. A necessary condition on the number of observations M required to find L inactive variables is given by

$$M \geq \max_{1 \leq j \leq K} \frac{\log \left[\binom{N-K+j}{j} \Big/ \binom{N-K+j-L}{j} \right]}{I(j)}.$$

That is, a lesser number of observations than the above will result in P_e being bounded strictly away from zero.

Finding Healthy Items via Non-adaptive Group Testing

- Noisy group testing signal model

$$\underline{y} = \bigvee_{i=1}^{N} \mathbf{D}_i \underline{x}_i \mathbb{I}_{\{i \in \mathcal{G}\}} \bigvee \underline{w}$$

  - $\mathcal{G}$ is the defective set
  - $\underline{x}_i \in \{0,1\}^M$ is the i^{th} column of $\mathbf{X}$
  - $\underline{w} \in \{0,1\}^M$ is the additive noise, $\underline{w}(i) \sim \mathcal{B}(q)$.
  - $\mathbf{D}_i \triangleq \text{diag}(\underline{d}_i)$
    - $\underline{d}_i \in \{0,1\}^M$, $\underline{d}_i(j) \sim \mathcal{B}(1-u)$ is chosen independently $\forall\, j = 1, 2, \ldots M$ and $\forall\, i = 1, 2, \ldots N$

Sufficient Number of Group Tests

	$0 \leq \alpha < 1$	Small α, e.g., $\alpha \leq 0.5$
No Noise	$O\left(\frac{K}{\log K}\frac{H_b(\alpha)}{(1-\alpha)}\right)$	$O\left(\frac{K\alpha}{\log K}\right)$
Dilution Noise	$O\left(\frac{K}{(1-u)\log K}\frac{H_b(\alpha)}{1-\alpha}\right)$	$O\left(\frac{K\alpha}{(1-u)\log K}\right)$
Additive Noise	$O\left(\frac{K}{\log\frac{1}{q}}\frac{H_b(\alpha)}{(1-\alpha)}\right)$	$O\left(\frac{K\alpha}{\log\frac{1}{q}}\right)$

- Different scenarios
  - Noiseless case: $u = 0, q = 0$
  - Additive noise model: $u = 0, q > 0$
  - Dilution noise model: $u > 0, q = 0$

Necessary Number of Group Tests

	$0 \leq \alpha < 1$	Small α, e.g., $\alpha \leq 0.5$
No Noise	$O\left(\frac{K}{\log K} \log \frac{1}{(1-\alpha)}\right)$	$O\left(\frac{K\alpha}{\log K}\right)$
Dilution Noise	$O\left(\frac{K}{(1-u)\log K} \log \frac{1}{1-\alpha}\right)$	$O\left(\frac{K\alpha}{(1-u)\log K}\right)$
Additive Noise	$O\left(\frac{K}{\log \frac{1}{q}} \log \frac{1}{(1-\alpha)}\right)$	$O\left(\frac{K\alpha}{\log \frac{1}{q}}\right)$

- Upper and lower bounds are order-wise tight

Sufficient number of tests: K grows linearly with N

Lemma

Let L_j and $C_2(L, N, K, j)$ be as defined before. Let $L < N - 2K$ and let $K \geq K_0$, where K_0 is some positive constant. Define $C_3 \triangleq -\log\left[1 - (1 - \frac{1}{K_0})^{K_0} + \exp(-2)\right]$. For the noiseless group testing case, if

$$M > \frac{1}{C_3} K \max_{1 \leq j \leq K} \frac{\log C_2(L, N, K, j)}{L_1} + \frac{\log K}{C_3},$$

then there exists a positive ϵ such that $P_e \leq \exp\left(-M\epsilon\right)$, and hence, $\lim_{N\to\infty} P_e = 0$.

Sufficient number of tests: K grows linearly with N

- Note that P_e is of the form $\exp(-M\epsilon)$ instead of $K\exp(-M\epsilon)$ previously
- Lower bound E_0 directly

$$E_0(\rho, 1, L_1) = -\log[(1-p)^{(K-1)}(1-p)^{(1+\rho L_1)} + p^{(1+\rho L_1)} + 1 - (1-p)^{(K-1)}]$$

- Derive the condition such that

$$ME_0(\rho, 1, L_1) - \rho \max_{1 \le j \le K} \log C_2(L, N, K, j) - \log K > 0$$

Conclusions

- Considered finding a subset of L healthy items in a large population containing K defective items
- Derived information theoretic bounds on the number of observations
- Contrasted two approaches:
  - Look in the complement of the set of sick items
  - Look for healthy items directly
- Impressive gains obtainable by directly identifying healthy items
- Specialized results to the nonadaptive group testing setup, accounting for additive noise and dilution

References

1 A. Sharma and C. R. Murthy, “On finding a subset of healthy individuals from a large population,” ArXiv Preprint, arXiv:1307.8240, Jul. 2013.
[Also presented at ITA 2013, San Diego, USA.]

2 G. Atia and V. Saligrama, “Boolean compressed sensing and noisy group testing,” IEEE Trans. Inf. Theory, vol. 58, no. 3, pp. 1880–1901, 2012

3 G. Atia and V. Saligrama, “A mutual information characterization for sparse signal processing,” in ICALP, Switzerland, 2011

4 R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, Inc., 1968

Thank You

Backup Slides

Sufficient number of tests: $K = 1$ case, Proof Sketch

- Gallager bounding technique
- Let X_1 be the active variable
- Sort $P(\underline{\mathbf{y}}|\mathbf{X}_i)$ for all $i = 1, 2, \ldots, N$
- The decoding algorithm will make an error if $P(\underline{\mathbf{y}}|\mathbf{X}_1)$ falls within the last L entries
- $\mathcal{E} \triangleq \{\text{error}|X_1 \text{ is active}, \mathbf{X}_1, \underline{\mathbf{y}}\}$

$$P_e = \sum_{\underline{\mathbf{y}}, \mathbf{X}_1} P(\underline{\mathbf{y}}|\mathbf{X}_1) Q(\mathbf{X}_1) \Pr(\mathcal{E}).$$

Sufficient number of tests: $K = 1$ case, Proof Sketch

- $S_z \subset [N] \backslash 1$ such that $|S_z| = N - L$
- $\mathcal{S}_\mathbf{z}$ denote a set of all possible S_z
- $\mathcal{A}_{S_z} = \{\mathbf{X}_{S_z} : P(\underline{\mathbf{y}}|\mathbf{X}_j) \; \geq \; P(\underline{\mathbf{y}}|\mathbf{X}_1) \, \forall \, j \in S_z\}$
- $\mathcal{E} \subset \mathcal{A} \triangleq \bigcup_{S_z \in \mathcal{S}_z} \mathcal{A}_{S_z}$,
    - An error event implies that there exists a set of $N - L$ variables, S_z, such that $P(\underline{\mathbf{y}}|\mathbf{X}_j) \; \geq \; P(\underline{\mathbf{y}}|\mathbf{X}_1) \, \forall \, j \in S_z$
- $\Pr(\mathcal{E}) \leq \Pr(\mathcal{A})$

Sufficient number of tests: $K = 1$ case, Proof Sketch

$$\Pr(\mathcal{E}) \leq \sum_{S_z \in \mathcal{S}_{\mathbf{z}}} \sum_{\mathbf{X}_{S_z} \in \mathcal{A}_{S_z}} Q(\mathbf{X}_{S_z})$$

$$\leq \sum_{S_z \in \mathcal{S}_{\mathbf{z}}} \sum_{\mathbf{X}_{S_z} \in \mathcal{A}_{S_z}} Q(\mathbf{X}_{S_z}) \prod_{j \in S_z} \left[\frac{P(\underline{\mathbf{y}}|\mathbf{X}_j)}{P(\underline{\mathbf{y}}|\mathbf{X}_1)} \right]^s$$

$$\leq \sum_{S_z \in \mathcal{S}_{\mathbf{z}}} \sum_{\mathbf{X}_{S_z}} \prod_{j \in S_z} Q(\mathbf{X}_j) \left[\frac{P(\underline{\mathbf{y}}|\mathbf{X}_j)}{P(\underline{\mathbf{y}}|\mathbf{X}_1)} \right]^s$$

$$= \binom{N-1}{L-1} \left\{ \sum_{\mathbf{X}_j} Q(\mathbf{X}_j) \left[\frac{P(\underline{\mathbf{y}}|\mathbf{X}_j)}{P(\underline{\mathbf{y}}|\mathbf{X}_1)} \right]^s \right\}^{N-L}$$

Sufficient number of tests: $K = 1$ case, Proof Sketch

- Let $0 \leq \rho \leq 1$
  - If the R.H.S. above is less than 1, then raising it to the power ρ makes it bigger, and if it is greater than 1, it remains greater than 1 after raising it to the power ρ
  - Thus,

$$\Pr(\mathcal{E}) \leq \binom{N-1}{L-1}^{\rho} \left\{ \sum_{\mathbf{X}_j} Q(\mathbf{X}_j) \left[\frac{P(\underline{\mathbf{y}}|\mathbf{X}_j)}{P(\underline{\mathbf{y}}|\mathbf{X}_1)} \right]^{s} \right\}^{\rho(N-L)}$$

Sufficient number of tests: $K = 1$ case, Proof Sketch

- Substitute back in the first error expression,

$$P_e \leq \binom{N-1}{L-1}^{\rho} \sum_{\underline{\mathbf{y}}} \sum_{\mathbf{X}_1} Q(\mathbf{X}_1) P(\underline{\mathbf{y}}|\mathbf{X}_1)^{1-\rho(N-L)s}$$

$$\left\{ \sum_{\mathbf{X}_j} Q(\mathbf{X}_j) P(\underline{\mathbf{y}}|\mathbf{X}_j)^{s} \right\}^{\rho(N-L)}$$

- Set $s = 1/(1+\rho(N-L))$
- Finally, use the independence across observations to obtain the final expression

Necessary Conditions: Proof Idea

- A genie-aided lower bound, using Fano's inequality
- Let E be the error event
- Let $S_\omega = S^{(j)} \cup S^{(K-j)}$, where $|S^{(j)}| = j$ and $|S^{(K-j)}| = K - j$ and $S^{(j)} \cap S^{(K-j)} = \{\emptyset\}$
- Consider $H(\omega, E|\underline{\mathbf{y}}, \mathbf{X}_{S^{(K-j)}})$

Necessary Conditions: Proof Idea

$$H(\omega, E|\underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) = H(E|\underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) + H(\omega|E, \underline{\mathbf{y}}, \mathbf{X}_{S(K-j)})$$

$$\leq H_b(P_e) + (1 - P_e)H(\omega|E = 0, \underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) + P_e H(\omega|E = 1, \underline{\mathbf{y}}, \mathbf{X}_{S(K-j)})$$

$$\leq H_b(P_e) + (1 - P_e)\log\binom{N-K+j-L}{j} + P_e \log\binom{N-K+j}{j}$$

Necessary Conditions: Proof Idea

$$\begin{aligned} H(\omega, E|\underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) &= H(\omega|\underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) + H(E|\omega, \underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) \\ &= H(\omega|\underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) \end{aligned}$$

- $H(\omega|\mathbf{X}_{S(K-j)}) = \log \binom{N-K+j}{j}$

$$\begin{aligned} \log \binom{N-K+j}{j} &= H(\mathbf{X}_{S_\omega}|\underline{\mathbf{y}}, \mathbf{X}_{S(K-j)}) + I(\mathbf{X}_{S_\omega}; \underline{\mathbf{y}}|\mathbf{X}_{S(K-j)}) \\ &\leq H_b(P_e) + \log \binom{N-K+j-L}{j} + \\ &\quad P_e\Gamma_I(L, N, K, j) + I(\mathbf{X}_{S_\omega}; \underline{\mathbf{y}}|\mathbf{X}_{S(K-j)}) \end{aligned}$$

Necessary Conditions: Proof Idea

- Noting that

$$I(\mathbf{X}_{S(j)}; \underline{\mathbf{y}} | \mathbf{X}_{S(K-j)}) \leq MI(X_{S(j)}; Y | X_{S(K-j)}) = MI^{(j)}.$$

- Lower bound on P_e is derived

$$P_e \geq 1 - \frac{H_b(P_e) + MI^{(j)}}{\Gamma_I(L, N, K, j)} \quad \forall\, j = 1, 2, \ldots, K$$
$$\geq 1 - \max_{1 \leq j \leq K} \frac{H_b(P_e) + MI^{(j)}}{\Gamma_I(L, N, K, j)}.$$