Semiquantitative Group Testing

Pradip Sasmal

IISc Bangalore

02.02.2019

Group testing

- m tests designed to identify d defectives among n items. Goal : minimize m
- probabilistic GT : a probability distribution is considered for d, and the goal is to minimize the expected number of tests
- combinatorial GT (CGT) : d (or at least an upper bound on d) is known in advance
- nonadaptive GT: all the tests are designed in advance
- adaptive GT: the result of one test may be used to govern the design of other tests

continue ....

- $C_{m \times n}$: binary test matrix, d_i : number of defectives participated in $i-$th test, y_i : $i-$th test output.
- quantitative GT (QGT): $y_i = d_i$
- the threshold group testing (TGT) model:

$$y_i = \begin{cases} 0 & \text{if } d_i < \eta_i \\ 1 & \text{if } d_i > \eta^i \\ 0 \text{ or } 1 & \text{if } \eta_i \leq d_i \leq \eta^i \end{cases}$$

η_i and η^i fixed lower and upper threshold respectively

Quantitative GT

- semiquantitative group testing (SQGT), motivated by a class of problems arising in genome screening experiments
- genotyping methods allow for more precise readings at the output than classical GT detectors, but still do not provide full information about the abundance of a target gene in the test
- codes constructed for CGT or TGT underutilize the potential of these sequencers, while codes constructed for QGT are prone to errors due to "overestimating" the sequencers' precision

SQGT vs other GTs

- test matrix $C_{m \times n}$: interger valued
- $c_j \in [q]^m$: codeword of $j-$th item
- $c_j(i)$: $i-$th entry of c_j : the amount of $j-$th sample used in the $i-$th test
- y_i : non-binary value that depends on the number of defectives through a given set of thresholds
- integer-valued test matrices as opposed to real-valued matrices : that the sample preparation in genotyping performed by robotic arms that are usually programmed to sample the same amount of DNA

Mathematical formulation

Definition

The "SQ-sum" of a set of $s \geq 1$ codewords, $\chi = \{x_1, x_2, \ldots, x_s\} = \{x_i\}_{i=1}^{s}$, in a SQGT model with thresholds $\eta = [\eta_1, \ldots, \eta_Q]^T$, is represented by $y_\chi = \odot_{i=1}^{s} x_j = x_1 \odot x_2 \odot \ldots_s$, and describes a vector of length m with its $k-$th coordinate equal to

$$y_\chi(k) = r \quad \text{if } \eta_r \leq \sum_{j=1}^{s} x_j(k) < \eta_{r+1},\, 0 \leq r \leq Q,$$

where $x_j(k)$ is the $k-$th coordinate of x_j , and $"+"$ stands for real-valued addition. $y_\chi \in [Q]^m$ is refereed as the syndrome of χ , and the underlying $\odot$ operation as the SQ-sum.

continue ..

SQGT model:

$$y = \odot_{j=1}^{d} x_{i_j},$$

x_{i_j} is the codeword of the $j-$th defective.

Example

Let $d = 3$, $m = 5$, $n = 10$, $q = 3$, $Q = 4$, and $\eta = [0, 2, 3, 5, 7]$.

$$\begin{bmatrix} 1 \\ 1 \\ 3 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 1 & 2 & 0 & 0 & 2 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\ 2 & 0 & 2 & 2 & 0 & 2 & 1 & 1 & 1 & 1 \\ 0 & 2 & 1 & 0 & 2 & 0 & 1 & 2 & 0 & 0 \\ 1 & 1 & 0 & 2 & 1 & 1 & 1 & 1 & 2 & 1 \end{bmatrix} \odot \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Special cases of SQGT

- if $q = Q = 2$ and $\eta_1 = 1$, SQRT $\implies$ CGT.
- if $Q - 1 = d(q-1)$ and $\forall\, r \in [Q]$, $\eta_r = r$, SQGT $\implies$ QGT, with possibly non-binary test matrix
- Assume $\eta_Q > (q-1)d$, SQGT with equidistant threshold : $\eta_r = r\eta$, where $r \in [Q+1]$ and $y_\chi(k) = \left\lfloor \frac{\sum_{j=1}^{s} x_j(k)}{\eta} \right\rfloor$

Definition

A set of codewords $\chi = \{x_i\}_{j=1}^{s}$ with syndrome y_χ is said to be included in another set of codewords $Z = \{z_i\}_{j=1}^{t}$ with syndrome y_Z, if $\in \{1, dots, m\}$, $y_\chi(i) \leq y_Z(i)$. Denote this inclusion property by $\chi \triangleleft Z$, or equivalently, $y_\chi \triangleleft y_Z$.

- Using this definition, it can be easily verified that if $\chi \subseteq Z$, then $\chi \triangleleft Z$.

SQ-disjunct code for error free SQGT

Definition

(SQ-disjunct code) : A code is called a $[q; Q; \eta; (1 : d); 0]-$SQ-disjunct code of length m and size n if $\forall s, t \leq d$ and for any sets of $q-$ary codewords $\chi = \{x_i\}_{j=1}^{s}$ and $Z = \{z_i\}_{j=1}^{t}$, $\chi \triangleleft Z$ implies $\chi \subseteq Z$.

Theorem

A $[q; Q; \eta; (1 : d); 0]-$SQ-disjunct code is capable of identifying any number of defectives less than or equal to d in the absence of test errors. In other words, given an error-free vector of test results $y \in [Q]^m$, any codeword with a syndrome included in y corresponds to a defective, and any codeword with a syndrome not included in y corresponds to a non-defective.

Theorem

A code is $[q; Q; \eta; (1 : d); 0]-$SQ-disjunct if and only if no codeword is included in a set of d other codewords.

SQ-disjunct code for SQGT with error

Definition

A code is called a $[q; Q; \eta; (1:d); e]-$SQ- disjunct code of length m and size n if for any set of $d+1$ codewords, $\chi = \{x_j\}_{j=1}^{d+1}$, and for any codeword $x_i \in \chi$, there exists a set of coordinates,R_i, of size at least $2e+1$ such that $\forall k_i \in R_i$,

$$y_{\{x_i\}}(k_i) > y_{\chi \backslash \{x_i\}}(k_i),$$

and R_i is disjoint of any R_ℓ for which $x_\ell \in \chi$ and $\ell \neq i$; in this equation $y_{\{x_i\}}$ is the syndrome of $\{x_i\}$, and $y_{\chi \backslash \{x_i\}}$ is the syndrome of the remaining d codewords in χ .

Decoding Algorithm

The decoding algorithm for a $[q; Q; \eta; (1:d); e]$—-SQ-disjunct code of length m and size n works as follows: For each codeword x_i, $i \in \{1, \ldots, n\}$, count the number of coordinates of $y_{\{x_i\}}$ for which

$$y_{\{x_i\}}(k) > y(k).$$

If the number of such coordinates is at least $e+1$, x_i does not correspond to a defective. On the other hand, if the number of such coordinates is at most e, the codeword corresponds to a defective.

- The computational complexity of the decoding algorithm is $O(mn)$.

Construction of SQ-disjunct codes

Construction 1 : Any code generated by multiplying a conventional binary $d-$disjunct code capable of correcting e errors by $q-1$, where $q-1 \geq \eta_1$, is a $[q; Q; \eta; (1:d); e]-$SQ-disjunct code.

Construction 2 : Form a matrix $C \in \{0, \eta, 2\eta, \ldots, I\eta\}^{m \times n}$ by choosing each entry independently according to the following probability distribution,

$$P_X(x) = \begin{cases} P_0 & \text{if } x = 0 \\ p_1 & \text{if } x \in \{0, \eta, 2\eta, \ldots, I\eta\} \end{cases}$$

where $I = \left\lfloor \frac{q-1}{\eta} \right\rfloor$, $P_0 = \frac{d}{d+1}$ and $P_0 = \frac{1}{I(d+1)}$. Then C is a $[q; Q; \eta; (1 : d); e]-$SQ-disjunct code of length m_I and size n_I with probability at least $1 - o(1)$; asymptotically, m_I equals

$$m_I \sim \frac{m_1}{\left(1 + \frac{1}{I^{d+1}d^d} \sum_{k=0}^{d-1} \binom{d}{k} \binom{I}{d-k+1} (ID)^k\right)}$$

where m_1 is the length of a $[q; Q; \eta; (1 : d); e]-$SQ-disjunct code of size $n_1 = n_I$, obtained by multiplying the best probabilistically constructed binary $d-$disjunct code, capable of correcting up to e errors, by η.

References

A. Emad and O. Milenkovic, “Semiquantitative Group Testing,” in IEEE Transactions on Information Theory, vol. 60, no. 8, pp. 4614-4636, Aug. 2014.