COMO POUPAR TESTES DE RASTREIO: A TESTAGEM EM GRUPOS COMO INTRODUÇÃO AO MÉTODO PROBABILÍSTICO

João Ribeiro
NOVA LINCS e FCT-UNL
jribeiro@tecnico.ulisboa.pt, joao.ribeiro@fct.unl.pt | https://sites.google.com/site/joaorib94/

O rastreio de doenças é uma peça fundamental no controlo de epidemias. Neste artigo estudamos um modelo matemático (simplificado) de testagem paralela. Veremos como o método probabilístico em combinatória nos permite facilmente obter esquemas de testagem com um número quase optimal de testes para uma dada população.

1. INTRODUÇÃO

Todos nós tivemos, recentemente, contacto próximo com testes de rastreio. Num contexto onde é necessário testar uma grande população num curto espaço de tempo e com recursos limitados, é fundamental correr o menor número possível de testes, e organizar o processo de testagem de forma altamente paralelizada. Este problema não é novo – foi estudado pela primeira vez há mais de 80 anos por Dorfman [Dor43], que pretendia otimizar o processo de testagem de soldados infetados com sífilis durante a Segunda Guerra Mundial.

Dorfman propôs o conceito de *pooling*: amostras recolhidas de vários soldados seriam misturadas e testadas em conjunto. Um teste positivo levaria, portanto, à conclusão de que pelo menos um dos soldados nesse grupo foi infetado (mas não *quantos* nem *quais* dos soldados). O objetivo seria usar o conceito de *pooling* para desenhar esquemas de testagem que, para uma população de n soldados, necessitassem de muito menos do que n testes.

No caso mais extremo, esquemas de testagem "em grupo" prosseguem em sequência. Isto é, o resultado dos primeiros i testes informa o conjunto de soldados a selecionar para o $(i+1)$-ésimo teste. Por exemplo, se sabemos que um teste que inclui dez soldados devolveu "negativo", então podemos removê-los da nossa população para os testes seguintes, pois sabemos que nenhum destes soldados está infetado. Apesar de poder levar a uma grande poupança no número de testes necessários para identificar os indivíduos infetados, estes esquemas em série requerem muito tempo e mão de obra, pois cada teste apenas pode ser determinado e efetuado quando os anteriores terminam. Consequentemente, existe um grande foco em desenhar esquemas de testagem em grupo que requerem um pequeno número de fases paralelas. O caso ideal corresponde a apenas uma fase de testagem, sendo portanto todos os testes definidos *a priori* e efetuados em paralelo. Chamamos a esquemas de testagem em grupo com esta propriedade *esquemas de testagem paralela*.

Neste artigo fazemos uma breve exploração da matemática por detrás dos esquemas de testagem paralela. É importante realçar que o modelo matemático considerado para estes testes, que herdamos de Dorfman, é simplificado e não se ajusta diretamente à realidade do rastreio de doenças infecciosas. Por exemplo, este modelo ignora, por um lado, efeitos de diluição de amostras e, por outro, ignora informação extra que pode ser revelada pelos testes. Também não consideramos questões de sensibilidade nem de especificidade (conceitos relacionados com a probabilidade de falsos negativos e falsos positivos) dos testes. Não obstante, ao longo das últimas décadas, o estudo de esquemas de testagem paralela estabeleceu-se como uma área madura com fortes ligações a combinatória, álgebra e teoria da probabilidade. Notavelmente, estes esquemas têm encontrado inúmeras aplicações a tópicos ortogonais à sua motivação original, como a transmissão de informação através de canais com ruído, aprendizagem automática, compressão de dados e sequenciação de ADN (incluindo o *Human Genome Project* [HG]).

A nossa exploração de esquemas de testagem paralela servirá como uma boa introdução ao método probabilístico. Esforçámo-nos por tornar a discussão acessível, necessitando o leitor apenas de conhecimentos básicos de combinatória e probabilidade discreta. Para não obscurecer as ideias mais interessantes com tecnicalidades, optámos por não otimizar certos argumentos. Discussões muito mais extensas e otimizadas destes tópicos podem ser encontradas nos livros de Du e Hwang [DH00, HD06] e na monografia de Aldridge, Johnson, e Scarlett [AJS19]. No que diz respeito a aplicações, [AJS19, Secção 1.7] apresenta uma discussão extensa e variada.

2. REPRESENTAÇÕES DE ESQUEMAS DE TESTAGEM PARALELA

Começamos por apresentar uma forma simples e útil de representar esquemas de testagem paralela. Dada uma população de n indivíduos, vemos cada teste como um vetor binário $v \in \{0,1\}^n$ onde $v_i = 1$ se e só se o item i participa no teste em causa. Por exemplo, se temos uma população de oito indivíduos e um dos testes é aplicado a uma mistura de amostras dos indivíduos 1, 2 e 5, então esse teste é representado pelo vetor $(1,1,0,0,1,0,0,0)$. Consequentemente, podemos representar um esquema de testagem paralela para uma população de n indivíduos com t testes como uma matriz binária M de dimensão $t \times n$ em que cada linha representa um teste.

Dado um vetor binário u, definimos o seu *suporte*, denotado por $\mathsf{supp}(u)$, como

$$\mathsf{supp}(u) = \{i \mid u_i = 1\}.$$

Por exemplo, se $u = (1,1,0,0,1,0,0,0)$ como acima, então $\mathsf{supp}(u) = \{1,2,5\}$, o que corresponde às coordenadas com valor 1.

Dizemos que um vetor binário u é *coberto* por outro vetor v com o mesmo comprimento se $\mathsf{supp}(u) \subseteq \mathsf{supp}(v)$. Por exemplo, o vetor $u = (1,1,0,0,1,0,0,0)$ é coberto pelo vetor $v = (1,1,1,0,1,0,1,0)$, pois para todas as coordenadas i tal que $u_i = 1$ também temos $v_i = 1$. Por outro lado, u não é coberto pelo vetor $w = (1,0,1,1,1,1,1,1)$, pois $u_2 = 1$ e $w_2 = 0$.

Podemos representar os indivíduos infetados numa população de n indivíduos como um vetor $x \in \{0,1\}^n$ em que $x_i = 1$ significa que o i-ésimo indivíduo está infetado. Portanto, o conjunto de indivíduos infetados corresponde a $\mathsf{supp}(x)$. A aplicação do esquema de testagem paralela definido por $M \in \{0,1\}^{t\times n}$ à população representada por x produz o vetor de resultados $y \in \{0,1\}^t$ tal que

$$\mathsf{supp}(y) = \bigcup_{j\in\mathsf{supp}(x)} \mathsf{supp}(M_{\cdot j}),$$

onde $M_{\cdot j}$ denota a j-ésima coluna de M. Dizemos que M é um *esquema de testagem paralela para n indivíduos e até d infetados* se para qualquer população $x \in \{0,1\}^n$ com $|\mathsf{supp}(x)| \leq d$ (i.e., a população x tem no máximo d infetados) conseguimos determinar $\mathsf{supp}(x)$ dada apenas a descrição M do esquema de testagem paralela e o vetor de resultados y. O esquema de testagem paralela mais básico corresponde ao caso em que M é a matriz identidade de dimensões $n \times n$.

Para facilitar a compreensão, apresentamos um exemplo de cálculo do vetor de resultados y a partir de uma matriz M e de um vetor de infetados x. Sejam

$$M = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

e $x = (1,1,0,1,0,0)$. Isto corresponde ao caso em que o primeiro, o segundo e o quarto indivíduos numa população estão infetados. Primeiro, verificamos que $\mathsf{supp}(x) = \{1,2,4\}$. Assim, sabemos que $y \in \{0,1\}^4$ é tal que $\mathsf{supp}(y) = \mathsf{supp}(M_{\cdot 1}) \cup \mathsf{supp}(M_{\cdot 2}) \cup \mathsf{supp}(M_{\cdot 4})$. Como

$$\mathsf{supp}(M_{\cdot 1}) = \{1,4\},\ \mathsf{supp}(M_{\cdot 2}) = \{2,4\},$$
$$\text{e } \mathsf{supp}(M_{\cdot 4}) = \{1,2\},$$

segue que

$$\mathsf{supp}(y) = \{1,4\} \cup \{2,4\} \cup \{1,2\} = \{1,2,4\},$$

e portanto $y = (1,1,0,1)$. Por outras palavras, o primeiro, o segundo e o quarto testes tiveram resultado positivo.

2.1. O Problema da Moeda Defeituosa

Iniciamos agora a nossa exploração de esquemas de testagem paralela, focando-nos no caso especial em que d=1, isto é, em que pretendemos apenas identificar até 1 infetado. Este caso corresponde ao seguinte conhecido enigma: temos um conjunto de n moedas com o mesmo peso, exceto no máximo uma delas, e temos também acesso a uma balança. De quantas pesagens necessitamos para identificar a moeda defeituosa, ou para provar que esta não existe?

Trivialmente, n pesagens são suficientes para identificar a moeda defeituosa. Veremos agora como a representação de esquemas de testagem paralela discutida acima nos permite facilmente resolver este enigma e mostrar que o número de testes realmente necessários é significativamente menor.

Teorema 1. *Existe um esquema de testagem paralela para uma população de tamanho n e até 1 infetado com $\lceil \log_2(n+1) \rceil$ testes, onde $\lceil \cdot \rceil$ denota o menor inteiro maior do que o argumento ou igual (o teto).*

Demonstração. Começamos por identificar as propriedades que tal esquema de testagem paralela terá. Seja $M \in \{0,1\}^{t\times n}$ uma qualquer matriz binária, e seja $x \in \{0,1\}^n$ o vetor de infetados com $|\mathsf{supp}(x)| \leq 1$. Recordamos que ao usarmos Mcomo esquema de testagem paralela, obtemos o vetor de resultados y tal que $\mathsf{supp}(y) = \bigcup_{j\in\mathsf{supp}(x)} \mathsf{supp}(M_{\cdot j})$.

Caso $|\mathsf{supp}(x)| = 0$ (i.e., não existem infetados), então $y = (0, \ldots, 0)$. Por outro lado, quando $\mathsf{supp}(x) = \{j\}$, então $y = M_{\cdot j}$. Concluímos que para obter o esquema de testagem paralela desejado basta construir uma matriz M tal que todas as colunas sejam distintas e nenhuma contenha apenas 0s. Isto pode ser feito representando os inteiros de 1 até n como sequências binárias de comprimento $\lceil \log_2(n+1) \rceil$ e usando estas sequências como colunas de M. □

Por exemplo, para identificarmos até $d = 1$ infetado numa população de $n = 7$ indivíduos basta usarmos $t = 3$ testes paralelos definidos pela matriz

$$M = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Se o vetor de infetados é $x = (0,0,0,0,1,0,0)$, então o correspondente vetor de resultados y satisfaz $y = M_{\cdot 5} = (1,0,1)$. No geral, se $\mathsf{supp}(x) = \{i\}$, então $y = M_{\cdot i}$. Como as colunas de M são todas distintas, conseguimos recuperar o indivíduo infetado i a partir de y. Quando $x = (0,0,0,0,0,0,0)$ (não há infetados), o vetor de resultados associado é $y = (0,0,0)$, que também é diferente de todas as colunas de M. Conseguimos, então, também discernir o caso em que não existem indivíduos infetados.

O Teorema 1 é otimal relativamente ao número de testes, o que pode ser demonstrado através de um simples argumento de contagem. Qualquer esquema de testagem paralela para n indivíduos que detete até 1 infetado tem de desambiguar entre $n+1$ eventos possíveis. Mais precisamente, tem de especificar que não existe nenhum infetado, ou, caso contrário, qual a posição do infetado entre os n indivíduos. Por outro lado, a aplicação de um esquema de testagem paralela com t testes gera um vetor binário $v \in \{0,1\}^t$, e, portanto, pode desambiguar entre no máximo 2^t casos possíveis. Concluímos, então, que é necessário ter

$$2^t \geq n+1.$$

Como t é inteiro, isto implica que $t \geq \lceil \log_2(n+1) \rceil$, que corresponde ao número de testes obtido no Teorema 1. Notavelmente, este argumento aplica-se igualmente ao caso em que os testes não são efetuados de forma paralela! Consequentemente, não perdemos nada em considerar apenas testes paralelos neste caso especial com no máximo $d = 1$ infetado.

3. MATRIZES DISJUNTAS

Na Secção 2.1 vimos um esquema simples para detetar até 1 infetado entre n indivíduos usando muito menos do que n testes. Gostaríamos, claro, de estender este resultado para o caso em que pretendemos detetar até $d > 1$ infetados de forma eficiente. Tendo isto em conta, definimos *matrizes disjuntas*, um conceito combinatorial simples e bastante útil no desenho de esquemas de testagem paralela com descodificação eficiente e que requerem poucos testes.

Definição 1 (Matriz d-disjunta) Uma matriz binária $M \in \{0,1\}^{t \times n}$ *diz-se* d-disjunta *se para quaisquer índices* $1 \leq j_1 < j_2 < \cdots < j_d \leq n$ *e* $j \notin \{j_1, \ldots, j_d\}$ *temos que*

$$\mathsf{supp}(M_{\cdot j}) \not\subseteq \mathsf{supp}(M_{\cdot j_1}) \cup \mathsf{supp}(M_{\cdot j_2}) \cup \cdots \cup \mathsf{supp}(M_{\cdot j_d}).$$

Por palavras, uma matriz M é d-disjunta se qualquer coluna de M não é coberta pela união (entrada a entrada) de quaisquer outras d colunas de M. Outra maneira equivalente de descrever esta propriedade é dizer que para quaisquer índices $1 \leq j_1 < j_2 < \cdots < j_d \leq n$ e $j \notin \{j_1, \ldots, j_d\}$ existe um índice i tal que $M_{ij} = 1$, mas $M_{ij_1} = M_{ij_2} = \cdots = M_{ij_d} = 0$. Por exemplo, a matriz

$$M = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

é 1-disjunta pois nenhuma coluna é coberta por outra coluna, mas não é 2-disjunta pois $\mathsf{supp}(M_{\cdot 1}) = \{1,2\}$ e $\mathsf{supp}(M_{\cdot 2}) \cup \mathsf{supp}(M_{\cdot 3}) = \{1,3\} \cup \{2,3\} = \{1,2,3\}$ e, portanto, $\mathsf{supp}(M_{\cdot 1}) \subseteq \mathsf{supp}(M_{\cdot 2}) \cup \mathsf{supp}(M_{\cdot 3})$. Já a matriz identidade

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

é 2-disjunta.

O seguinte teorema captura a utilidade das matrizes disjuntas.

Teorema 2. *Se uma matriz* $M \in \{0,1\}^{t \times n}$ *é* d*-disjunta, então representa um esquema de testagem paralela para* n *indivíduos e até* d *infetados.*

Demonstração. Sejam $M \in \{0,1\}^{t \times n}$ uma matriz d-disjunta, $x \in \{0,1\}^n$ tal que $|\mathsf{supp}(x)| \leq d$, e $y \in \{0,1\}^t$ o respetivo vetor de resultados. Recordamos que y é tal que $\mathsf{supp}(y) = \bigcup_{j \in \mathsf{supp}(x)} \mathsf{supp}(M_{\cdot j})$. Provamos que o conjunto de indivíduos infetados, $\mathsf{supp}(x)$, corresponde

exatamente aos índices $j \in \{1, \ldots, n\}$ tal que a coluna $M_{\cdot j}$ é coberta por y.

Primeiro, notamos que se j é um indivíduo infetado, então y claramente cobre $M_{\cdot j}$, pois $\mathsf{supp}(M_{\cdot j}) \subseteq \mathsf{supp}(y)$ pela definição de y. Suponhamos agora que j não está infetado. Isto quer dizer que $j \notin \mathsf{supp}(x)$. Como $|\mathsf{supp}(x)| \leq d$ e M é d-disjunta, concluímos que

$$\mathsf{supp}(M_{\cdot j}) \not\subseteq \bigcup_{j' \in \mathsf{supp}(x)} \mathsf{supp}(M_{\cdot j'}) = \mathsf{supp}(y),$$

e, portanto, $M_{\cdot j}$ não é coberta por y. □

Uma propriedade importante implícita na demonstração do Teorema 2 é que esquemas de testagem paralela provenientes de matrizes disjuntas têm um simples algoritmo associado de reconstrução do conjunto $\mathsf{supp}(x)$ de infetados – basta verificar quais as colunas de M que são cobertas pelo vetor de resultados y. No geral, não se conhecem algoritmos de reconstrução eficientes para esquemas de testagem paralela para até d infetados cujas matrizes associadas não são d-disjuntas. Para exemplificar o algoritmo de reconstrução para matrizes disjuntas, considere-se a matriz

$$M = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

que é 2-disjunta. Pelo Teorema 2, concluímos que conseguimos identificar até 2 infetados com a ajuda do esquema de testagem descrito por M (com um número sub otimal de testes!). Suponhamos que recebemos o vetor de testes $y = (1, 1, 1, 0, 1, 0)$. Para recuperar o conjunto de infetados $\mathsf{supp}(x)$, procuramos as colunas de M que são cobertas por y. Como apenas $M_{\cdot 1}$ e $M_{\cdot 3}$ são cobertas por y, concluímos que os indivíduos 1 e 3 estão infetados.

Note-se que se M é um esquema de testagem paralela para n indivíduos e até d infetados, não é necessariamente verdade que M é também d-disjunta. Basta considerar a matriz M descrita na Secção 2.1, que não é 1-disjunta mas permite identificar até $d = 1$ infetado. No entanto, estas duas noções são "quase" equivalentes – o leitor curioso poderá convencer-se de que tal matriz M é sempre $(d-1)$-disjunta.

4. LIMITES DE ESQUEMAS DE TESTAGEM PARALELA PARA MAIS INFETADOS

Tendo em conta o Teorema 1, sabemos como detetar 1 infetado entre n indivíduos usando apenas $\lceil \log_2(n+1) \rceil$ testes. Vimos também que este número de testes é otimal através de um simples argumento de contagem. Considerando agora $d > 1$ infetados, quantos testes serão necessários?

Claro, n testes serão sempre suficientes, mas esperamos usar muito menos testes. Podemos facilmente generalizar o argumento de contagem para d infetados. Observemos que existem mais de

$$\binom{n}{\leq d} := \sum_{i=0}^{d} \binom{n}{i}$$

eventos possíveis, correspondentes aos possíveis subconjuntos de indivíduos infetados. Isto implica que necessitamos de

$$t \geq \log_2 \binom{n}{\leq d} \geq d \log_2(n/d) \tag{1}$$

testes para desambiguar entre estes casos, onde usámos a conhecida desigualdade $\binom{n}{\leq d} \geq \binom{n}{d} \geq (n/d)^d$, válida para todo o d tal que $1 \leq d \leq n$. Existem, no entanto, melhores minorantes para o número de testes necessários. Por exemplo, Füredi [Für96] apresenta uma prova elegante de que

$$t \geq \frac{1}{4} d^2 \log_d n \tag{2}$$

quando $d \leq n^\alpha$ para qualquer $\alpha < 1/2$ e $n \geq n_0$ para uma certa constante n_0. Anteriormente, sabíamos já por D'yachkov e Rykov [DR82] que é possível substituir a constante $1/4$ por uma expressão mais complicada que é aproximadamente igual a $1/2$ quando d é pequeno. De facto, até hoje este continua a ser o melhor minorante conhecido num extenso leque de parâmetros d e n. A exemplo de comparação, quando $d = n^{1/4}$, o minorante dado pela Equação (1) é da ordem $n^{1/4} \log_2 n$, enquanto o minorante da Equação (2) é da ordem $\sqrt{n}$, levando assim a uma grande diferença assintótica.

5. ESQUEMAS DE TESTAGEM PARALELA E O MÉTODO PROBABILÍSTICO

Será que o minorante dado pela Equação (2) está muito longe da verdade? Para respondermos a esta questão, usaremos o *método probabilístico*. Esta estratégia elegante foi ini-

cialmente usada por Szele [Sze43] e Erdős [Erd47] em teoria de grafos e por Shannon [Sha48a, Sha48b] em teoria da informação. Para estabelecermos a existência de objetos (por exemplo, matrizes binárias $t \times n$) com uma dada propriedade desejada (por exemplo, ser d-disjunta), começamos por definir uma experiência aleatória apropriada sobre um conjunto de objetos. De seguida, demonstramos que esta experiência aleatória devolve um objeto com a propriedade desejada com probabilidade positiva. Isto garante a existência do objeto desejado de forma "não construtiva".

Desde a sua introdução, o método probabilístico tem encontrado inúmeras aplicações em matemática discreta e na teoria da computação. O livro de Alon e Spencer [AS16] é uma excelente fonte de tais aplicações. A aplicação do método probabilístico ao estudo de esquemas de testagem paralela que discutiremos serve como boa introdução a esta técnica.

O seguinte teorema estabelece, com a ajuda do método probabilístico, a existência de um esquema de testagem paralela para n indivíduos e até d infetados com aproximadamente $4d^2 \ln n$ testes. Ao contrário dos exemplos básicos mais comuns da aplicação do método probabilístico, que impõem distribuições uniformes na classe de objetos e cuja respetiva análise é diretamente substituível por um simples argumento de contagem, o exemplo que veremos nesta secção usa uma distribuição não uniforme sobre o espaço das matrizes binárias $M \in \{0,1\}^{t\times n}$ que depende do parâmetro d. É, portanto, mais aparente a grande utilidade do método probabilístico nesta situação.

Teorema 3. *Existe uma matriz d-disjunta $M \in \{0,1\}^{t\times n}$ com $t = \lceil e\, d(d+1)\ln n\rceil$. Em particular, existe um esquema de testagem paralela para n indivíduos e até d infetados com $t = \lceil e\, d(d+1)\ln n\rceil$ testes.*

Demonstração. Aplicamos o método probabilístico. Construímos uma matriz binária $M \in \{0,1\}^{t\times n}$ decidindo que cada entrada é, independentemente das restantes, 1 com probabilidade $p \in [0,1]$ (e 0 caso contrário), onde o número de testes t e a probabilidade p são parâmetros que escolheremos mais tarde.

Fixemos quaisquer $d+1$ colunas de M. A probabilidade de a primeira coluna ser coberta pela união das restantes d colunas é exatamente

$$(1-p(1-p)^d)^t. \tag{3}$$

Para nos convencermos disto, começamos por observar que a primeira entrada da primeira coluna não é coberta pela união das primeiras entradas das restantes d colunas exatamente quando a primeira entrada da primeira coluna é 1 e as restantes primeiras entradas são todas 0. Isto acontece com probabilidade $p(1-p)^d$. Para que a primeira coluna seja coberta pela união das restantes, este evento não pode ocorrer em nenhuma das t coordenadas. Obtemos então a Equação (3) usando a independência entre as várias coordenadas.

A matriz M é d-disjunta se e só se o evento acima não ocorrer para nenhuma escolha de $d+1$ colunas de M e seleção da primeira coluna. Existem n maneiras de escolher a "coluna-alvo", e, posteriormente, $\binom{n-1}{d}$ maneiras de escolher d colunas dentre as restantes $n-1$ colunas de M. Para cada uma destas escolhas a probabilidade da coluna alvo ser coberta pela união das restantes é dada pela Equação (3). Usando o facto de que

$$\Pr[E_1 \vee E_2 \vee \cdots \vee E_m] \leq \sum_{i=1}^{m} \Pr[E_i]$$

para quaisquer eventos $E_1, \ldots, E_m$ (a famosa *union bound* ou desigualdade de Boole), temos que a matriz M não é d -disjunta com probabilidade menor ou igual a

$$n \cdot \binom{n-1}{d} \cdot (1-p(1-p)^d)^t.$$

Falta apenas, então, escolher p e t apropriadamente para que tenhamos

$$n \cdot \binom{n-1}{d} \cdot (1-p(1-p)^d)^t < 1, \tag{4}$$

caso em que concluímos que existe uma matriz d-disjunta $M \in \{0,1\}^{t\times n}$. Começamos por notar que

$$(1-p(1-p)^d)^t \leq e^{-tp(1-p)^d},$$

consequência da desigualdade $(1+x)^t \leq e^{tx}$ válida para todo o $x \geq -1$ e $t \geq 0$. Como

$$n \cdot \binom{n-1}{d} \leq n(n-1)^d,$$

onde usamos a desigualdade $\binom{a}{b} \leq a^b$ válida para todos os naturais $a \geq b$, para a equação (4) se verificar é então suficiente escolhermos p e t tal que

$$n \cdot (n-1)^d \cdot e^{-tp(1-p)^d} < 1. \tag{5}$$

Começamos por analisar o termo $p(1-p)^d$ a fim de otimizarmos a escolha de p. Com $d \geq 1$ fixo, a função $f(p) = p(1-p)^d$ é maximizada em $]0,1[$ quando $p = \frac{1}{d+1}$. Isto acontece pois

$$f'(p) = (1-p)^d - dp(1-p)^{d-1}$$

e, portanto $f'(p) \geq 0$ exatamente quando $p \leq \frac{1}{d+1}$, com igualdade quando $p = \frac{1}{d+1}$. Usando $p = \frac{1}{d+1}$, caso em que

$$f(p) = \tfrac{1}{d+1} \cdot \left(\tfrac{d}{d+1}\right)^d,$$

ao aplicarmos logaritmos a ambos os lados da Equação (5) concluímos que t satisfaz a Equação (5) desde que

$$t > \left(\frac{d+1}{d}\right)^d (d+1)(d\ln(n-1) + \ln n) =$$

$$= (1+1/d)^d (d+1)(d\ln(n-1) + \ln n).$$

Como $(1+1/d)^d < e$ para todo o $d \geq 1$ e $d\ln(n-1) + \ln n \leq (d+1)\ln n$, concluímos que é suficiente termos

$$t \geq e\, d(d+1)\ln n.$$

Podemos escolher, então, $t = \lceil e\, d(d+1)\ln n \rceil$, como desejado.

Comparando o Teorema 3 com o minorante da Equação (2), vemos que estes diferem apenas num fator multiplicativo de ordem $\ln d$. Concluímos que os minorantes e majorantes aqui discutidos estão perto do número ótimo de testes, e que, quando d é suficientemente pequeno comparado com n, o número de testes necessário é também muito mais pequeno do que o majorante trivial, n. Com vista a uma comparação mais concreta, a Tabela 1 apresenta o número de testes suficientes para detetar até d infetados numa população de n indivíduos para várias escolhas destes parâmetros. □

d \ n	10^2	10^3	10^4	10^5	10^6
2	76	113	151	188	226
5	376	564	752	939	1127
10	1377	2066	2754	3443	4131
20	5278	7887	10506	13145	15773

Tabela 1: Número de testes suficientes pelo Teorema 3 para detetar até d infetados numa população de n indivíduos, para várias escolhas de (d, n). Entradas a vermelho representam os casos em que o número de testes é maior do que n, o majorante trivial.

Finalmente, é interessante também realçar que a demonstração do Teorema 3 garante não só a existência de um esquema de testagem paralela com poucos testes, mas também especifica um simples e prático algoritmo probabilístico que recebe como *input* naturais n e $d \leq n$ e devolve com probabilidade de, pelo menos, 0.99 um esquema de testagem paralela para n indivíduos e até d infetados com um número ligeiramente mais elevado de testes $t = \lceil 20d(d+1)\ln n \rceil$.

6. INDO MAIS ALÉM

A nossa exploração de esquemas de testagem paralela foi, necessariamente, superficial. Por exemplo, apesar de o Teorema 3 garantir a existência de esquemas de testagem paralela que usam poucos testes, este não providencia um algoritmo *determinístico e eficiente* que, dados o tamanho da população n e o número máximo de infetados como *inputs*, devolva uma matriz d-disjunta $M \in \{0,1\}^{t\times n}$. O desenho e análise de tais algoritmos foi alvo de grandes esforços. Um resultado clássico de Kautz e Singleton [KS64] apresenta um algoritmo eficiente que, dados n e d, constrói matrizes d-disjuntas $M \in \{0,1\}^{t\times n}$ com $t = cd^2\log_d^2 n$ para uma constante $c > 0$ independente de n e d. Quando d é muito mais pequeno do que n, este valor de t é assintoticamente maior do que o valor de t obtido no Teorema 3. Por outro lado, quando, por exemplo, $d = n^\alpha$ para alguma constante $\alpha > 0$, o número de testes do método de Kautz-Singleton é muito mais pequeno assintoticamente do que o número de testes garantido pelo Teorema 3. Esta linha de investigação culminou também num algoritmo eficiente de Porat e Rothschild [PR08] que, dados n e d, constrói matrizes d-disjuntas $M \in \{0,1\}^{t\times n}$ com $t = Cd^2\log_2 n$ para uma constante $C > 0$ grande. Este valor de t é da ordem do valor não construtivo apresentado no Teorema 3, mas não é necessariamente prático devido à constante C acima.

O estudo moderno de esquemas de testagem paralela também considera outras direções motivadas por aplicações práticas. Por exemplo, podemos permitir uma pequena probabilidade de erro na recuperação do conjunto de indivíduos infetados, sendo esta probabilidade tomada sobre uma escolha uniformemente aleatória do conjunto de d infetados e, se relevante, sobre a aleatoriedade usada para construir o esquema de testagem. Neste caso, usando uma estratégia semelhante à da demonstração do Teorema 3, é possível concluir que o número de testes suficientes é da ordem de $d\log_2(n/d)$ [AJS19, Secções 1.3 e 1.4 e Capítulo 2]. Uma longa sequência de trabalhos tem também vindo a estudar esquemas de testagem resilientes a erros nos resultados dos testes [AJS19, Capítulo 3]. Os esquemas apresentados neste breve artigo não têm garantias de resiliência, e garantem a reconstrução correta de qualquer conjunto de até d infetados. Numa direção ortogonal, vários grupos

têm estudado esquemas de testagem paralela que permitem a reconstrução extremamente rápida do conjunto de infetados. Como ponto de partida, o algoritmo de reconstrução associado a esquemas de testagem paralela baseados em matrizes disjuntas $M \in \{0,1\}^{t \times n}$ tem complexidade temporal de ordem $n \cdot t$. Atualmente, conhecemos construções de esquemas de testagem paralela com um número quase-otimal de testes e com algoritmos de reconstrução associados com complexidade temporal quase-ótima (na ordem de $t \log_2^2 n$) [CN20].

Os esquemas estudados neste artigo englobam apenas uma fase de testagem que pode ser completamente paralelizada. Outra direção de investigação recai sobre esquemas que consistem em múltiplas fases de testagem sequenciais, o que leva a uma ainda maior redução do número de testes à custa de limites adicionais na paralelização. Exemplos particulares são o esquema original de Dorfman [Dor43] e esquemas baseados em *square arrays* [PS94].

Agradecimento: O autor agradece ao revisor anónimo por vários comentários e sugestões pertinentes que melhoraram a exposição deste artigo.

REFERÊNCIAS

[AJS19] Matthew Aldridge, Oliver Johnson and Jonathan Scarlett. "Group Testing: An Information Theory Perspective". *Foundations and Trends in Communications and Information Theory*, 15(3-4):196–392, 2019.

[AS16] Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley & Sons, 2016.

[CN20] Mahdi Cheraghchi and Vasileios Nakos. "Combinatorial Group Testing and Sparse Recovery Schemes with Near-Optimal Decoding Time". In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science* (FOCS), pages 1203–1213, 2020.

[DH00] Ding-Zhu Du and Frank Hwang. *Combinatorial Group Testing and its Applications*. World Scientific, 2000.

[Dor43] Robert Dorfman. "The Detection of Defective Members of Large Populations". *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[DR82] Arkadii Georgievich D'yachkov and Vladimir Vasil'evich Rykov. "Bounds on the Length of Disjunctive Codes". *Problemy Peredachi Informatsii*, 18(3):7–13, 1982.

[Erd47] Paul Erdős. "Some remarks on the theory of graphs". *Bulletin of the American Mathematical Society*, 53(4):292–294, 1947.

[Für96] Zoltán Füredi. "On r-cover-free families". *Journal of Combinatorial Theory, Series A*, 73(1):172–173, 1996.

[HD06] Frank Hwang and Ding-Zhu Du. *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing*. World Scientific, 2006.

[HG] The Human Genome Project. *https://www.genome.gov/human-genome-project*.

[KS64] William Kautz and Richard Singleton. "Nonrandom binary superimposed codes". *IEEE Transactions on Information Theory*, 10(4):363–377, 1964.

[PR08] Ely Porat and Amir Rothschild. "Explicit non-adaptive combinatorial group testing schemes." In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfsdóttir, and Igor Walukiewicz, editors, *Automata, Languages and Programming*, pages 748–759. Springer Berlin Heidelberg, 2008.

[PS94] Ravindra Phatarfod and Aidan Sudbury. "The use of a square array scheme in blood testing". *Statistics in Medicine*, 13(22):2337–2343, 1994.

[Sha48a] Claude Shannon. "A Mathematical Theory of Communication". *Bell System Technical Journal*, 27(3):379–423, 1948.

[Sha48b] Claude Shannon. "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27(4):623–656, 1948.

[Sze43] Tibor Szele. "Kombinatorikai vizsgalatok az iranyitott teljes graffal kapcsolatban". *Középiskolai Matematikaiés Fizikai Lapok*, 50:223–256, 1943.

SOBRE O AUTOR

João Ribeiro é atualmente Professor Auxiliar no Departamento de Informática da FCT-UNL e membro integrado do NOVA LINCS. Em agosto de 2024 inicia funções como Professor Auxiliar no Departamento de Matemática do IST. Antes disto, saltitou entre Portugal, Suíça, Reino Unido, e EUA. Os seus interesses centram-se na teoria da computação, com ênfase em criptografia, pseudoaleatoriedade, e teoria de códigos e da informação.