Modelos Estocásticos e Modelos Espaciais – 6ª Feira, 21 de Abril, Auditório A (10h40)

Deteção de distribuições atípicas em sequências genómicas

Ana Helena Tavares[1], Vera Afreixo[2], Paula Brito[3]
[1] *Departamento de Matemática & CIDMA, Universidade de Aveiro, ahtavares@ua.pt;*
[2] *Departamento de Matemática & CIDMA & iBiMED, Universidade de Aveiro; vera@ua.pt;*
[3] *FEP & LIAAD-INESC TEC, Universidade do Porto, mpbrito@fep.up.pt*

Sumário: Neste trabalho abordamos o problema de deteção de distribuições atípicas no contexto das distâncias entre palavras genómicas, seguindo uma abordagem funcional, tendo em conta não apenas a existência de outliers de magnitude, mas também a possível existência de outliers de forma. Aplica-se um procedimento baseado na medida *directional outlyingness*, que permite detetar *outliers* de magnitude e *outliers* de forma.

Palavras-chave: *Directional outlyingness*, Distribuição atípica, Palavra genómica.

O DNA pode ser descrito por uma longa sequência de A, C, G e T's. Apesar de utilizar um alfabeto de apenas 4 letras, é nesse texto que se encontra codificada a informação necessária à vida dos organismos. A deteção de palavras genómicas atípicas, ou *outliers*, é um assunto de muito interesse no estudo do DNA. Neste trabalho utilizamos as distâncias entre palavras como descritor da distribuição de uma palavra ao longo do genoma.

A distância entre-palavras é definida como a diferença entre a última letra de duas ocorrências sucessivas da mesma palavra. Por exemplo, na sequência ACTCGACAC as distâncias entre os AC são 5 e 2. Olhando para um conjunto de distribuições de distâncias como um conjunto funcional, podemos utilizar medidas e técnicas de análise de dados funcionais, para a deteção de *outliers*. Neste contexto, uma função *outlier* não é apenas aquela que contém valores atipicamente baixos ou elevados (*outliers* de magnitude), mas também aquela que apresente um padrão diferente do da maioria (*outliers* de forma).

Recentemente, Rousseeuw *et. al.* (2016) propuseram uma nova abordagem, baseada na medida *directional outlyingness* (DO), que atribui um valor robusto de atipicidade a cada ponto do domínio de uma função. Baseados na medida de *outlyingness* de Stahel-Donoho (Stahel, 1981; Donoho, 1982) definem a medida DO de um ponto y, em relação a uma amostra univariada $Y = \{y_1, \cdots, y_m\}$, como

$$DO(y;Y) = \begin{cases} \dfrac{y - Med(Y)}{S_a(Y)}, & se\ y \geq Med(Y) \\ \dfrac{y - Med(Y)}{S_b(Y)}, & se\ y \leq Med(Y) \end{cases}$$

onde S_a e S_b são estimadores de escala robustos para os subconjuntos de pontos acima e abaixo da mediana, respetivamente. Considere-se agora uma função f, um conjunto de funções $F = \{f_1, \dots, f_m\}$, todas de domínio univariado, e $\{t_1, \dots, t_T\}$ um conjunto discreto do domínio onde o valor das funções é observado. Para cada ponto t_i pode calcular-se o DO

de $f(t_i)$ com respeito ao conjunto de valores tomados pelas outras funções nesse mesmo ponto. Os valores de DO obtidos para cada função podem ser reduzidos a um par ordenado, por via de uma medida de localização, *fDO*, (uma média ponderada dos valores DO) e uma medida de variabilidade, *vDO*. Representando os pares ordenados (*fDO*, *vDO*) num gráfico de dispersão obtém-se o *functional outlyingness map* (FOM), uma ferramenta gráfica que permite identificar possíveis candidatos a *outliers*.

Neste trabalho segue-se essa abordagem na deteção de palavras *outliers* no genoma humano. A cada conjunto funcional, F_k, formado pelas 4^k distribuições de distâncias entre palavras associadas às palavras de tamanho k, aplica-se o procedimento baseado na medida DO e na representação gráfica FOM ($k = 1, \dots, 5$). Numa primeira abordagem considera-se que as distâncias contribuem todas com o mesmo peso na definição do *fDO* (distribuição uniforme).

O FOM associado ao conjunto de distribuições F_2 evidencia a distribuição de distâncias entre-CG como *outlier*, pois esta apresenta valores elevados de *fDO* e de *vDO*. De facto, a distribuição de distâncias entre-CG distingue-se das restantes por ser mais "achatada". À medida que o valor de k aumenta, mais inesperado é o padrão de algumas curvas, apresentando um ou mais picos que podem ocorrer tanto em distâncias curtas como em distâncias mais longas. No caso de $k = 5$, o FOM resultante aponta 17 casos como *outliers*. O método permite capturar distribuições com um padrão que se diferencia do padrão da maioria, assim como distribuições com picos em sub-intervalos onde as outras distribuições não os apresentam. São efetuadas experiências utilizando diferentes função peso e os resultados são comparados.

Os resultados iniciais indicam que o procedimento baseado na medida DO é promissor, colocando em evidência distribuições atípicas que eram mascaradas pelas restantes curvas.

Agradecimentos: Este trabalho é parcialmente financiado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) e FCT (Fundação Portuguesa para a Ciência e Tecnologia) através dos projetos UID/MAT/04106/2013 do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações), UID/BIM/04501/2013 do Instituto de Biomedicina (iBiMED), UID/EEA/50014/2013 do INESC-TEC (Instituto de Engenharia de Sistemas e Computadores do Porto), POCI-01-0145-FEDER-006961 do COMPETE 2020 (Programa Operacional Competitividade e Internacionalização) e bolsa de doutoramento PD/BD/105729/2014 de AT.

Referências

Rousseeuw, Peter J., Raymaekers, J. & Hubert, M. (2016). A measure of directional outlyingness with applications to image data and video. *Preprint arXiv*:1608.05012.

Stahel, Werner A. (1981), Breakdown of Covariance Estimators, *Research Report*, vol. 31, Fachgruppe für Statistik, E.T.H. Zürich, Switzerland.

Donoho, David L. (1982), Breakdown properties of multivariate location estimators. Ph.D. diss., Harvard University.