II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Classificação hierárquica com distribuições *a posteriori*: Um estudo de simulação em padrões temporais de VIH

Diana Rocha[1], Sónia Gouveia[1,2], Carla Pinto[3,4], Manuel Scotto[5], João Nuno Tavares[3]

[1] Centro de Investigação e Desenvolvimento em Matemática e Aplicações - CIDMA, Univ de Aveiro
[2] Instituto de Engenharia Electrónica e Informática de Aveiro - IEETA, Universidade de Aveiro
[3] Centro de Matemática da Universidade do Porto - CMUP, Universidade do Porto
[4] Instituto Superior de Engenharia do Instituto Politécnico do Porto - ISEP
[5] CEMAT e Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa

RESUMO

O objetivo deste trabalho é avaliar diferentes abordagens para identificação de grupos de pacientes VIH com padrões temporais de evolução da doença similares. Foi considerado um sistema de equações diferenciais ordinárias para caracterizar a comportamento ao longo do tempo de um paciente VIH sob tratamento antiretroviral - TAR de longo prazo, com 5 parâmetros estimados a partir de metodologia Bayesiana. As distribuições *a posteriori* foram usadas para quantificar distâncias (univariadas) entre pacientes, através do valor médio da distribuição *a posteriori*, e considerando a distância entre as distribuições *a posteriori* para cada parâmetro. O resultado do agrupamento hierárquico obtido pelas duas abordagens sugere que o uso de uma distância que considere a distribuição *a posteriori* é preferível. Trabalho futuro irá considerar distâncias multivariadas em vez de distâncias univariadas.

Palavras e frases chave: Vírus da Imunodeficiência Humana (VIH), Estimação Bayesiana, Modelos Matemáticos, Classificação Hierárquica.

1. INTRODUÇÃO

Não existem estudos em Portugal que permitam a identificação de grupos de pacientes portadores de VIH, com padrões temporais de evolução da doença semelhantes. A identificação de grupos, além de ter relevância ao nível da epidemiologia da doença, pode permitir também definir terapêuticas mais ajustadas a cada (grupo de) paciente(s) semelhantes e assim conseguir atender simultaneamente à individualização do tratamento e à gestão mais eficiente dos recursos hospitalares. Por este motivo, há todo o interesse em desenvolver uma metodologia que permita fazer a identificação de grupos de pacientes com trajetórias similares. Neste trabalho considerou-se o seguinte sistema EDO (equações diferenciais ordinárias) para caracterizar a trajetória de cada paciente portador de VIH sob tratamentos TAR de longo prazo [2]

$$\begin{aligned} \frac{dT_U(t)}{dt} &= \lambda - \rho T_U(t) - \eta(t) T_U(t) V(t) \\ \frac{dT_I(t)}{dt} &= \eta(t) T_U(t) V(t) - \delta T_I(t) \\ \frac{dV(t)}{dt} &= N\delta T_I(t) - cV(t) \end{aligned} \tag{1}$$

com variáveis de estado $CD4 \equiv T(t) = T_U + T_I$, respetivamente a concentração de células CD4 não infetadas e infetadas, e V como a carga viral. Os parâmetros do sistema são λ (taxa de proliferação das células CD4), ρ e δ (respetivamente taxa de mortalidade de células CD4 não infetadas e infetadas), $\eta(t)$ (taxa de infeção, que varia ao longo do tempo e depende da eficácia do TAR), N (número médio de virions produzidos por célula infetada) e c (taxa de depuração de viriões livres). Este trabalho foi desenvolvido em 3 etapas. Primeiro, foram geradas 20 réplicas de dois padrões distintos de evolução temporal de $T(t)$ e $V(t)$ [2]. Segundo, para cada réplica, foram estimados os parâmetros do modelo. Finalmente, foram identificadas classes de pacientes por agrupamento hierárquico, tendo em conta que seria esperado existirem duas classes do dendrograma, cada uma das classes constituída por um dos padrões de evolução temporal considerado. O agrupamento hierárquico foi considerado com duas métricas de distância: para cada parâmetro do modelo, consideraram-se distâncias sobre o valor médio e distâncias baseadas na distribuição *a posteriori*. Os resultados das diferentes abordagens foram finalmente comparados por intermédio da dispersão intra e entre grupos.

2. MÉTODOS EXPERIMENTAIS

Neste trabalho foram geradas observações igualmente espaçadas das séries temporais $T(t) = T_U(t)+T_I(t)$ e $V(t)$, por resolução numérica do sistema 1 no intervalo de tempo $[0, 20]$(dias) através do método de Runge-Kutta. Foram consideradas as condições iniciais $T_U(0) = 600$, $T_I(0) = 30$, $V(0) = 105$, $\eta(t) = 9 * 10^{-5}(1 - 0{,}9\cos(\pi t/1000))$ e valores de $(\lambda, \rho, \delta, N, c)$ correspondentes a [2]. Foram também considerados dois conjuntos distintos para os parâmetros $(\lambda, \rho, \delta, N, c)$ referenciados na Tabela 1, os quais representam curvas que correspondem ao comportamento de indivíduos reais, sejam indivíduo 1 e 2 [2]. No final deste processo foram constituídos dois conjuntos de 19 observações não igualmente espaçadas no intervalo de tempo $[0, 20]$(dias), treze no primeiro dia de tratamento (internamento) e uma medição a cada dois dias, do dia 2 ao dia 14 do tratamento (ambulatório), um para cada indivíduo. Neste trabalho foram consideradas 10 réplicas de cada indivíduo, cada réplica foi constituída adicionando ao conjunto original de 19 observações um erro de 19 observações de CD4 disponíveis, assumindo uma distribuição Normal de média nula e variância ajustada de acordo com o facto de que as medições laboratoriais de CD4 estão sujeitas a um erro de aproximadamente 20 % em relação ao valor medido [3]. Para cada uma das 20 réplicas geradas, os parâmetros $(\lambda, \rho, \delta, N, c)$ foram estimados por intermédio do algoritmo Metropolis-Hastings com condições iniciais iguais às usadas para a geração dos dados. Consideraram-se as distribuições *a priori* Gamma, Normal e Wishart [1].

As métricas de distância univariadas para avaliar semelhança entre pacientes foram obtidas a partir das distribuições *a posteriori*. Para cada parâmetro, foi considerada a distância Euclideana entre a média do parâmetro obtida para a réplica i e para a réplica j, isto é a) $d^2_{ij} = (\bar{x_i} - \bar{x_j})^2$, e também a semelhança entre a distribuição acumulada *a posteriori* de cada parâmetro da réplica i e da réplica j, sejam $F_i(x)$ e $F_j(x)$, dada por b) $d^2_{ij} = \int_0^1 (F_i^{-1}(y) - F_j^{-1}(y))^2 y(1-y)dy \simeq \sum_{k=1}^{99} (\hat{F}_i^{-1}(y_k) - \hat{F}_j^{-1}(y_k))^2 y_k(1-y_k)$ com $y_k = k/100$. Para esta última distância são considerados diferentes pesos para cada percentil $F^{-1}(y)$, onde diferenças na mediana terão maior peso $(y = 0{,}5)$. Esta distância é estimada por intermédio da aproximação do integral por um somatório ponderado de áreas de trapézios, obtidos numa escala com 99 pontos igualmente espaçados num intervalo de $[0; 1]$.

A avaliação das métricas de distância para cada par de réplicas $i, j = 1, 2, \ldots, 20$ originou uma matriz 20×20 utilizada para obter um dendrograma por agrupamento/linkage sobre o valor médio do grupo. As abordagens de classificação foram comparadas pela variabilidade estimada intra e entre grupos.

3. ANÁLISE DE RESULTADOS

A Tabela 1 apresenta medidas descritivas sobre as distribuições *a posteriori* para cada parâmetro e réplica. Foi possível observar que a ordem de grandeza do erro de medição considerado não degradou de forma evidente o desempenho na estimação dos parâmetros do modelo.

Parâmetro	**Valor**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$log_{10}(\lambda)$	1,56	1,52 (2,41)	1,55 (2,38)	1,57 (2,40)	1,55 (2,40)	1,55 (2,41)	1,55 (2,43)	1,56 (2,40)	1,57 (2,41)	1,54 (2,43)	1,53 (2,41)
$log_{10}(\rho)$	−0,97	−0,97 (2,55)	−0,98 (2,57)	−0,95 (2,55)	−0,99 (2,57)	−0,94 (2,57)	−0,98 (2,56)	−0,94 (2,56)	−0,97 (2,54)	−1,00 (2,59)	−0,97 (2,55)
$log_{10}(\delta)$	−0,30	−0,29 (2,71)	−0,30 (2,72)	−0,30 (2,69)	−0,30 (2,71)	−0,30 (2,71)	−0,31 (2,71)	−0,30 (2,70)	−0,30 (2,72)	−0,30 (2,71)	−0,30 (2,68)
$log_{10}(c)$	0,48	0,45 (2,69)	0,49 (2,68)	0,46 (2,72)	0,50 (2,72)	0,49 (2,74)	0,50 (2,71)	0,48 (2,69)	0,46 (2,73)	0,49 (2,70)	0,49 (2,73)
$log_{10}(N)$	3,00	3,00 (2,55)	3,02 (2,55)	2,98 (2,56)	2,98 (2,57)	3,01 (2,56)	3,00 (2,55)	3,01 (2,58)	2,97 (2,56)	3,00 (2,55)	2,99 (2,57)
Parâmetro	**Valor**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
$log_{10}(\lambda)$	1,26	1,25 (2,41)	1,26 (2,40)	1,27 (2,42)	1,25 (2,41)	1,27 (2,42)	1,28 (2,39)	1,27 (2,39)	1,24 (2,43)	1,26 (2,41)	1,28 (2,41)
$log_{10}(\rho)$	−1,40	−1,38 (2,56)	−1,41 (2,57)	−1,37 (2,57)	−1,39 (2,55)	−1,40 (2,55)	−1,40 (2,56)	−1,40 (2,55)	−1,42 (2,59)	−1,39 (2,57)	−1,40 (2,54)
$log_{10}(\delta)$	−0,37	−0,39 (2,67)	−0,36 (2,72)	−0,40 (2,71)	−0,37 (2,70)	−0,38 (2,74)	−0,33 (2,72)	−0,34 (2,69)	−0,39 (2,70)	−0,39 (2,71)	−0,36 (2,68)
$log_{10}(c)$	0,58	0,54 (2,70)	0,56 (2,71)	0,59 (2,72)	0,58 (2,72)	0,59 (2,75)	0,59 (2,68)	0,58 (2,70)	0,57 (2,71)	0,59 (2,72)	0,57 (2,71)
$log_{10}(N)$	3,44	3,44 (2,58)	3,44 (2,56)	3,43 (2,57)	3,46 (2,59)	3,43 (2,59)	3,43 (2,55)	3,45 (2,55)	3,46 (2,56)	3,43 (2,54)	3,45 (2,56)

Tabela 1: Estimativas dos parâmetros do modelo EDO, média (desvio padrão), para as realizações do indivíduo 1 (1 a 10) e para as realizações do indivíduo 2 (11 a 20).

A Figura 1 apresenta os dendrogramas obtidos pelas métricas de distância a) e b) considerando cada um dos parâmetros do modelo. Para ambas as métricas, os parâmetros λ, ρ, N e c permitem genericamente uma discriminação adequada das 20 réplicas em 2 grupos distintos (réplicas de 1 a 10 do indivíduo 1 e réplicas de 11 a 20 do indivíduo 2). No entanto, para o caso do parâmetro δ, observa-se que na discriminação pela métrica a) existem dois elementos que se encontram no grupo que não o esperado (réplicas 16 e 17 no grupo das réplicas do indivíduo 1). Com a métrica b), apenas o elemento 16 se encontra no grupo que não o esperado. Como observado na tabela 1, estas classificações incorretas devem-se ao facto dos valores médios de δ para as réplicas 16 e 17 serem os mais baixos em relação aos restantes valores médios do mesmo grupo e mais próximos dos valores médios das réplicas do outro grupo. Assim sendo, os resultados sugerem que a métrica b) poderá menos sensível a grandes variações nas médias dos parâmetros.

Para comparar as métricas de distância a) e b), de forma mais quantitativa, foram calculados os coeficientes de variação para os valores da distância obtidos entre pares de réplicas. Para avaliar a variabilidade intra grupo, consideraram-se as distâncias entre réplicas da mesma classe, enquanto que para quantificar a variabilidade entre grupos, consideraram-se as distâncias entre réplicas de classes diferentes. Os resultados obtidos encontram-se na Tabela 2. Como é possível observar, a dispersão intra-grupo é menor para a métrica b), indicando que a métrica b) permite quantificar uma distância mais pequena quando as réplicas são efetivamente do mesmo grupo. A métrica intra-grupos foi igual ou ligeiramente menor para a métrica b), indicando uma pequena degradação na diferenciação dos dois grupos distintos sem, no entanto, alterar o desempenho na classificação.

4. CONCLUSÕES

Este trabalho considerou o agrupamento hierárquico de trajetórias temporais através de duas métricas univariadas: a) uma baseada nas diferenças entre as médias das distribuições *a posteriori* e b) outra baseada nas diferenças ponderadas entre os percentis das distribuições *a posteriori*, para cada parâmetro do modelo. Pelos resultados obtidos podemos verificar que em ambas as abordagens existe uma separação correta dos dois grupos de indíviduos (1 a 10 e de 11 a 20), excepto para o

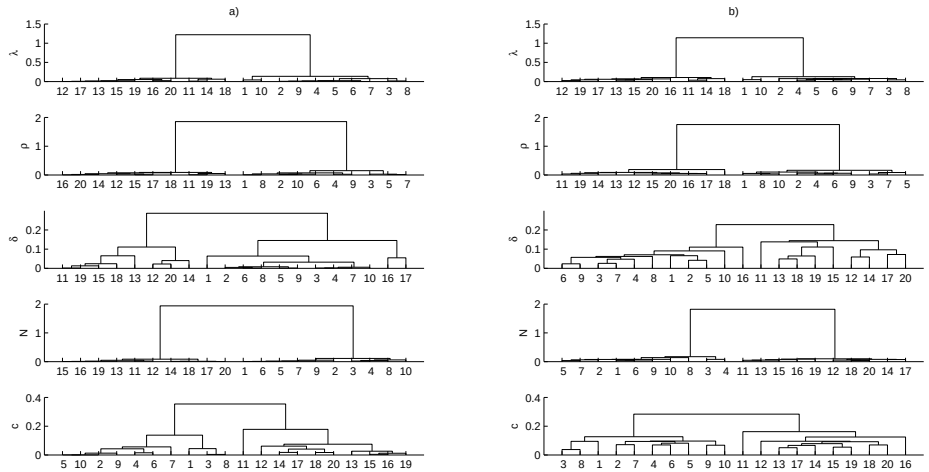

Figura 1: Dendrogramas obtidos pelas métricas de distância a) e b). Para mais informação, ver texto.

Parâmetro	**Intra-grupo (1 a 10)**		**Intra-grupo (11 a 20)**		**Entre-grupos**	
	a)	**b)**	**a)**	**b)**	**a)**	**b)**
λ	0.76	0.25	0.70	0.27	0.07	0.07
ρ	0.67	0.38	0.65	0.43	0.05	0.05
δ	0.82	0.26	0.68	0.33	0.32	0.26
c	0.73	0.21	0.77	0.25	0.25	0.20
N	0.68	0.41	0.68	0.27	0.04	0.04

Tabela 2: Coeficientes de variação intra e entre grupos, quantificados através dos valores da distância entre duas réplicas, respectivamente, da mesma classe e de classes diferentes. Para mais informação, ver texto.

parâmetro δ. No entanto, é também sugestivo de que uma métrica de distância multivariada, capaz de combinar adequadamente a informação distribucional de todos os parâmetros, poderá ainda melhorar o desempenho de classificação bem como diminuir a variabilidade intra-grupo.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e a Tecnologia, FCT, através de fundos nacionais (MEC) e estruturais europeus (FEDER), no âmbito dos projetos UID/MAT/04106/2013 (CIDMA/UA), UID/CEC/00127/2013 (IEETA/UA) e UID/MAT/00144/2013 (CMUP/UP). Diana Rocha agradece o financiamento concedido através da bolsa de Doutoramento FCT (ref. SFRH/BD/107889/2015).

Referências

[1] Huang, Y., Liu, D., Wu, H. (2004). Hierarchical bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* 62, 413–423.

[2] Liang, H., Miao, H., Wu, H. (2010). Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model. *The Annals of Applied Statistics* 4, 460–483.

[3] Whitby, L., Whitby, A.,Fletcher, M.,Helbert, M.,Reilly, J. T., Barnett, D. (2013). Comparison of Methodological Data Measurement Limits in CD41 T Lymphocyte Flow Cytometric Enumeration and Their Clinical Impact on HIV Management. *Cytometry Part B (Clinical Cytometry)* 84B, 248–254.