Análise Bayesiana Espacial Conjunta de Dados Longitudinais e de Sobrevivência com Aplicação ao Estudo do VIH/SIDA no Brasil

Rui Martins
Escola Superior de Saúde Egas Moniz e Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), *ruimartins@ymail.com*

Giovani L. Silva
CEAUL e Dep. Matemática - Instituto Superior Técnico, Universidade de Lisboa, *gsilva@math.ist.utl.pt*

Valeska Andreozzi
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), *valeska.andreozzi@fc.ul.pt*

Palavras–chave: Análise conjunta, Dados longitudinais, Dados de sobrevivência, Estatística espacial, Inferência bayesiana.

Resumo: A análise conjunta dos dados longitudinais e de sobrevivência tem suscitado muito interesse nos últimos anos, especialmente para dados relativos ao VIH/SIDA. Tradicionalmente, estes dados eram analisados considerando separadamente o tempo-até-evento (análise de sobrevivência) e as medições repetidas de biomarcadores (análise longitudinal). Mas, tendo em conta que os dados são observados num mesmo indivíduo, a modelação conjunta das respostas subjacentes parece ser mais apropriada. O modelo conjunto espacial bayesiano aqui proposto apresenta melhorias consideráveis nas distribuições do tempo mediano de sobrevivência quando comparado com o modelo separado.

1 Introdução

A análise conjunta de dados longitudinais e de sobrevivência tem em linha de conta qualquer associação que possa existir entre as medidas repetidas e o tempo até à ocorrência do evento de interesse. Tradicionalmente, estes dados eram modelados separadamente, mesmo quando existia, pelo menos, uma relação latente entre ambos. A construção do modelo conjunto começa, habitualmente, por considerar dois submodelos separados (longitudinal e de sobrevivência) que depois são ligados entre si.
Neste trabalho vamos centrar-nos no tempo-até-evento, sendo que a modelação em simultâneo das medidas repetidas de um biomarcador (considerado preditor da sobrevivência, *e.g.*, a contagem de linfócitos T $CD4^+$, doravante

designado simplesmente por CD4) permitirá que as estimativas dos parâmetros do modelo de sobrevivência reflictam correctamente a incerteza nas medições desse biomarcador, além de permitirem recuperar parte da informação perdida devido à censura.
Optou-se por uma abordagem completamente bayesiana, o que permitirá incorporar, de forma simples, efeitos aleatórios espaciais e, portanto, capturar heterogeneidades partilhadas, mas não observadas nos indivíduos que vivem numa mesma região. Tsiatis e Davidian [11], Ibrahim et al. [3] ou Rizopoulos [5, 6] são trabalhos que podem ser consultados para mais informações sobre esta temática.
O resto deste trabalho está organizado da seguinte forma. Na secção 2 descrevemos o conjunto de dados. Na secção 3 é desenvolvido o modelo espacial conjunto. Na secção 4 é realizada uma análise detalhada do conjunto de dados de VIH/SIDA aplicando os métodos propostos. A secção 5 discute a avaliação do modelo através de uma análise de resíduos de imputação múltipla e de uma análise de sensibilidade.

2 Descrição da estrutura dos dados

Os dados foram recolhidos numa rede de 88 laboratórios localizados nos 27 estados do Brasil, durante os anos de 2002-2006. As respostas recolhidas foram a contagem de células CD4 (uma medida do estado imunológico; valores mais elevados são indicadores de um indivíduo mais saudável) e o tempo de sobrevivência numa amostra aleatória de $n = 4653$ indivíduos. As variáveis explicativas são a idade (< 50 e ≥ 50 anos), sexo, IOPrev (indicador da presença de infecções oportunistas prévias), estado de residência, data de diagnóstico e data de morte (disponível se a morte ocorreu antes de 31 de Dezembro de 2006 e censurada caso contrário). O tempo de sobrevivência após o diagnóstico é o período de tempo entre a data de diagnóstico e data da morte. A variável idade foi estruturada com base em recomendações do Ministério da Saúde, pois o grupo com mais de 50 anos de idade apresenta maior proporção de atraso no início do tratamento quando comparado com o grupo populacional 15-49 anos (Souza et al. [9]).
As contagens de células CD4 por sexo, idade e IOPrev evidenciaram uma grande assimetria em direcção aos valores mais altos de CD4, sugerindo assim uma transformação (no caso usou-se a raiz quadrada). Houve 320 mortes, 88% dos pacientes tinham entre 15 e 49 anos; 2774 (60%) eram do sexo masculino, dos quais 220 morreram. 61% dos indivíduos não tinham infecções prévias; 6.7% vive no Centro-Oeste, 11.5 % no Nordeste, 4.8% no Norte, 60% no Sudeste e 16.7% no sul. A mediana de CD4 inicial foi de 245 células/mm^3 (homens - 226 células/mm^3; mulheres - 263 células/mm^3). Em média, os pacientes fizeram 4.62 exames de CD4, resultando num total de 21.508 observações.

3 Modelo conjunto

Em estudos ligados ao VIH/SIDA a contagem de CD4 dos pacientes é medida longitudinalmente e serve como biomarcador para o tempo até à morte do paciente. Consideremos que $y_{ik}(t_{ikj})$ denota a raiz quadrada do valor de CD4 na j-ésima medida repetida observada no instante t_{ikj} relativamente ao paciente i residente no k-ésimo estado do Brasil, $k = 1,\ldots,K$, $i = 1,\ldots,n_k$ e $j = 1,\ldots,m_{ik}$, onde n_k representa o número de pacientes que vivem no estado brasileiro k e m_{ik} o número de medidas repetidas de $\sqrt{\text{CD4}}$ do indivíduo ik (paciente i a viver no estado brasileiro k). O total de indivíduos é dado por $N = \sum_{k=1}^{K} n_k = 4653$ e o total de medidas repetidas no estudo é dado por $M = \sum_{k=1}^{K}\sum_{i=1}^{n_k} m_{ik} = 21508$. O vector $\mathbf{y}_{ik} = (y_{ik1},\ldots,y_{ikm})$ representa as m medidas repetidas observadas no indivíduo ik. Sejam T_{ik} o tempo de sobrevivência (possivelmente censurado à direita) para o paciente i do estado brasileiro k e δ_{ik} o seu indicador de ocorrência do evento ($\delta_{ik} = 1$, se o indivíduo morreu por causas ligadas à SIDA; $\delta_{ik} = 0$, se o indivíduo se mantinha vivo no final de 2006). Note-se que a *verdadeira* trajectória da resposta longitudinal nos instantes de medida, $\{y^*_{ik}(t_{ikj}), k = 1,\ldots,K, j = 1,\ldots,n_{ik}\}$, não é observada. De facto, o que se observa é a *verdadeira* trajectória da resposta longitudinal, $y^*_{ik}(t_{ikj})$, acrescida de algum erro. A verdadeira trajectória, sendo encarada como um factor latente, pode ser vista como representando o verdadeiro estado de saúde do indivíduo. Finalmente, representemos os dados observados para o indivíduo ik por $\mathcal{D}_{ik} = \{T_{ik},\delta_{ik},\mathbf{y}_{ik}\}$ que se supõe serem independentes entre os indivíduos, reflectindo a crença de que o processo da doença se desenvolve independentemente para cada sujeito.

3.1 Modelo longitudinal

Considere-se que o seguinte modelo linear com efeitos aleatórios descreve o processo longitudinal,

$$y_{ik}(t_{ikj}) = y^*_{ik}(t_{ikj}) + e_{ik}(t_{ikj}) = \mathbf{x}_{1ik}^\top(t_{ikj})\boldsymbol{\beta}_1 + W_{1ik}(t_{ikj}) + e_{ik}(t_{ikj}), \quad (1)$$

onde $e_{ik}(t_{ikj}) \sim \mathcal{N}(0,\sigma^2)$ representa um erro de medida, $\mathbf{x}_{1ik}^\top(t_{ikj})\boldsymbol{\beta}_1$ é a resposta média do indivíduo ik no instante t_{ikj} modelada em termos de um conjunto de variáveis explicativas $\mathbf{x}_{1ik}(t_{ikj})$ em cada data de exame de CD4, $W_{1ik}(t_{ikj}) = \mathbf{z}_{1ik}^\top(t_{ikj})\,\boldsymbol{b}_{ik}$ representa os efeitos aleatórios longitudinais, $\mathbf{z}_{1ik}^\top(t_{ikj})$ é a matriz de delineamento dos efeitos aleatórios e $\boldsymbol{b}_{ik}$ é o vector dos efeitos aleatórios individuais. Vamos recorrer à suposição habitual de que $\boldsymbol{b}_{ik}$ são independentes e identicamente distribuídos com $\mathcal{N}_p(\mathbf{0},\boldsymbol{\Sigma})$, sendo independente de $e_{ik}(t_{ikj})$. Para a resposta longitudinal $y_{ik}(t_{ikj})$, condicionada nos efeitos aleatórios $\boldsymbol{b}_{ik}$, suporemos uma distribuição Normal, ou seja, $y_{ik}(t_{ikj})|\boldsymbol{b}_{ik} \sim \mathcal{N}(y^*_{ik}(t_{ikj}),\sigma^2)$, evidenciando, assim, a independência condicional entre os diversos indivíduos.

3.2 Modelo de sobrevivência espacial

A associação entre o processo longitudinal e o processo de sobrevivência será feita através da partilha dos efeitos aleatórios individuais (Henderson et al. [2]), o que conduz a que os desvios individuais relativamente à trajectória global subjacente forneçam a informação necessária para a função de risco. Assim, o risco de morte será representado por

$$h_{ik}(t|\boldsymbol{b}_{ik},\mathbf{x}_{2ik}) = h_0(t)\exp\{\mathbf{x}_{2ik}^\top\boldsymbol{\beta}_2 + \boldsymbol{\gamma}W_{2ik}(t) + Q_k\}. \tag{2}$$

onde $\boldsymbol{\beta}_2$ é um vector de parâmetros que liga um vector de covariáveis de base $\mathbf{x}_{2ik}$ (podem ou não coincidir com $\mathbf{x}_{1ik}$) que contribuem para o risco, $W_{2ik}(t)$ é uma função dos efeitos aleatórios, $\boldsymbol{b}_{ik}$, e $\boldsymbol{\gamma}$ é o parâmetro de associação. Q_k representa um efeito aleatório espacial específico do estado brasileiro k e que captura o (log) risco relativo não explicado para um evento no estado k. O vector dos efeitos aleatórios, $\boldsymbol{b}_{ik}$, explica, não só a associação entre os dois processos, como também a correlação existente entre as medidas repetidas num mesmo sujeito. O processo latente W_{2ik} pode apresentar várias formas. Em particular, se o modelo longitudinal for de declive e ordenada na origem aleatórios, *i.e.*, se $W_{1ik}(t) = b_{1ik} + b_{2ik}(t)$, podemos considerar que $\boldsymbol{\gamma}W_{2ik}(t) = \gamma_1 b_{1ik} + \gamma_2 b_{2ik}$ onde $\boldsymbol{b}_{ik} = (b_{1ik},b_{2ik})$ e $\boldsymbol{\gamma} = (\gamma_1,\gamma_2)$ mede o efeito do número de células de CD4 no risco de morte, isto é, γ_1 e γ_2 medem a associação induzida através da ordenada na origem e do declive, respectivamente. Esta especificação permite que diferentes pacientes tenham medidas iniciais de CD4 distintas (aquando do diagnóstico) e também tendências temporais de contagens de CD4 diferentes durante o período de 2002-2006. Supõe-se ainda que o tempo-até-evento, T_{ik}, e o vector das medidas repetidas, $\mathbf{y}_{ik}$, são independentes dados os efeitos aleatórios individuais, $\boldsymbol{b}_{ik}$, e as covariáveis de interesse.

3.3 Verosimilhança

A contribuição do ik-ésimo indivíduo para a função de verosimilhança do processo longitudinal será denotada por L_{1ik} e para a verosimilhança relativa ao tempo de sobrevivência será denotada por L_{2ik}. A sua contribuição para a verosimilhança conjunta será então

$$L_{ik}(\Omega_{ik}|\mathcal{D}_{ik}) = L_{1ik}(\boldsymbol{b}_{ik},\boldsymbol{\beta}_1,\sigma^2|\mathbf{y}_{ik})L_{2ik}(\boldsymbol{b}_{ik},\boldsymbol{\beta}_2,\boldsymbol{\gamma},Q_k|T_{ik},\delta_{ik}) \tag{3}$$

onde Ω_{ik} denota a colecção de todos os parâmetros do modelo e as verosimilhanças $L_{1ik}(\boldsymbol{b}_{ik},\boldsymbol{\beta}_1,\sigma^2|\mathbf{y}_{ik})$ e $L_{2ik}(\boldsymbol{b}_{ik},\boldsymbol{\beta}_2,\boldsymbol{\gamma},Q_k|T_{ik},\delta_{ik})$ são, respectivamente,

$$\prod_{j=1}^{m_{ik}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\left[y_{ik}(t_{ikj}) - \mathbf{x}_{1ik}^\top(t_{ikj})\boldsymbol{\beta}_1 - \mathbf{z}_{1ik}^\top(t_{ikj})\boldsymbol{b}_{ik}\right]^2}{2\sigma^2}\right\} \tag{4}$$

$$S(t_{ik})^{1-\delta_{ik}} \times \left(-\frac{d}{dt_{ik}} S(t_{ik})\right)^{\delta_{ik}} = h(t_{ik})^{\delta_{ik}} S(t_{ik}), \tag{5}$$

onde $S(t_{ik}) = P(T \geq t_{ik}|\boldsymbol{b}_{ik},\boldsymbol{\beta}_2,\boldsymbol{\gamma},Q_k)$. No caso de T_{ik} ter uma distribuição Exponencial, i.e., $\mathcal{E}(e^{\eta_{ik}})$, temos que $h(t_{ik}) = \exp\{\eta_{ik}\}$ e $S(t_{ik}) = \exp\{-e^{\eta_{ik}} t_{ik}\}$, sendo que $\eta_{ik} = \mathbf{x}_{2ik}^{\top}\boldsymbol{\beta}_2 + \boldsymbol{\gamma} W_{2ik}(t) + Q_k$. Portanto, a verosimilhança conjunta é simplesmente $L(\boldsymbol{\Omega}|\mathcal{D}) = \prod_{k=1}^{K}\prod_{i=1}^{n_k} L_{ik}(\Omega_{ik}|\mathcal{D}_{ik})$.

4 Aplicação

A trajectória longitudinal específica de cada indivíduo ik será modelada através do modelo misto (1), ou seja,

$$\begin{aligned} y_{ik}(t_{ikj}) = {} & \beta_{11} + \beta_{12} t_{ikj} + \beta_{13}\texttt{sexo}_{ik} + \beta_{14}\texttt{idade}_{ik} + \beta_{15}\texttt{IOPrev}_{ik} + \\ & b_{1ik} + b_{2ik} t_{ikj} + e_{ikj}. \end{aligned} \tag{6}$$

Para os tempos de sobrevivência considerámos a distribuição Weibull, i.e., $T_{ik} \sim \mathcal{W}(\rho,e^{\eta_{ik}(t)})$, e um seu caso particular quando $\rho = 1$ (distribuição Exponencial), onde

$$\eta_{ik}(t) = \beta_{21} + \beta_{22}\texttt{sexo}_{ik} + \beta_{23}\texttt{idade}_{ik} + \beta_{24}\texttt{IOPrev}_{ik} + \gamma_1 b_{1ik} + \gamma_2 b_{2ik} + Q_k. \tag{7}$$

As estimativas *a posteriori* dos parâmetros foram obtidas através de simulações por métodos de Monte Carlo via cadeias de Markov (MCMC) implementados no software WinBUGS [4], usando distribuições *a priori* vagas, mas próprias, para os parâmetros do modelo. Denotando as distribuições Gama, Normal, Wishart e Uniforme respectivamente por $\mathcal{G}$, $\mathcal{N}$, $\mathcal{W}ish$ e $\mathcal{U}$, as especificações *a priori* são $1/\sigma^2 \sim \mathcal{G}(0.01,0.01)$, $\boldsymbol{\beta}_1 = (\beta_{11},\beta_{12},\beta_{13},\beta_{14},\beta_{15}) \sim \mathcal{N}_5(\mathbf{0}_5,1000I_5)$, $\boldsymbol{\beta}_2 = (\beta_{21},\beta_{22},\beta_{23},\beta_{24}) \sim \mathcal{N}_4(\mathbf{0}_4,1000I_4)$, $\boldsymbol{b}_{ik} = (b_{1ik},b_{2ik}) \sim \mathcal{N}_2(\mathbf{0},\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}ish(1000I_2,\xi)$, $\gamma_1 \sim \mathcal{N}(0,100)$, $\gamma_2 \sim \mathcal{N}(0,100)$. Supusemos um modelo auto-regressivo condicional intrínseco (ICAR) para os efeitos aleatórios espaciais, i.e., $Q_k|\sigma_Q^2 \sim ICAR(\sigma_Q^2)$, sendo que $1/\sigma_Q^2 \sim \mathcal{G}(0.5, 0.0005)$, e que todos os parâmetros eram independentes *a priori*.
Durante a análise foram ajustados uma série de diferentes modelos, sendo que a selecção destes foi baseada no valor da medida DIC (*Deviance Information Criterion*) [10] (vide Tabela 1). Os parâmetros de regressão das covariáveis `sexo`, `idade` e `IOPrev` são sempre incluídos nos dois submodelos. De referir que os modelos XI e XII apresentam a possibilidade de existirem efeitos aleatórios espaciais estruturados (Q_k) e não estruturados ($s_k|\sigma_s^2 \overset{iid}{\sim} \mathcal{N}(0,\sigma_s)$, $k = 1,\ldots,K$), isto é, que não consideram a estrutura de vizinhanças dos diversos estados brasileiros presente nos dados.

4.1 Análise dos resultados

O menor valor do DIC foi obtido com o modelo IX, pelo que foi este modelo o modelo seleccionado. A média *a posteriori* para o parâmetro de forma, ρ, da

Model	W_1	W_2	Q	p_D	DIC
I	0	0	0	9.855	139004
II	b_1	0	0	4050.21	119686
III	b_1	$\gamma_1 b_1$	0	4064.08	119359
IV	$b_1 + b_2 t$	0	0	6425.54	112251
V	$b_1 + b_2 t$	$\gamma_1 b_1$	0	6424.89	112026
VI	$b_1 + b_2 t$	$\gamma_2 b_2$	0	6445.97	112228
VII	$b_1 + b_2 t$	$\gamma(b_1 + b_2)$	0	6434.78	111948
VIII	$b_1 + b_2 t$	$\gamma_1 b_1 + \gamma_2 b_2$	0	6461.27	111907
IX	$b_1 + b_2 t$	$\gamma_1 b_1 + \gamma_2 b_2$	Q_k	6458.72	**111891**
X	$b_1 + b_2 t$	$\gamma(b_1 + b_2)$	Q_k	6437.54	111952
XI	$b_1 + b_2 t$	$\gamma(b_1 + b_2)$	$Q_k + s_k$	6464.34	111903
XII	$b_1 + b_2 t$	$\gamma(b_1 + b_2)$	s_k	6464.55	111899

Tabela 1: Comparação de modelos bayesianos.

distribuição Weibull foi de 1.04. Dado que este valor não é muito diferente de 1 e devido ao elevado tempo de computação (cerca de 25 horas), decidimos considerar apenas os casos em que $\rho = 1$, isto é, um modelo exponencial. Os nossos resultados foram baseados em 100000 iterações com um período de aquecimento de 20000 iterações. Para obviar os problemas de autocorrelação dentro das cadeias utilizou-se um espaçamento de 50 resultando num total de 1600 iterações utilizadas para obtenção dos valores *a posteriori*. Não foram detectados problemas de convergência com as amostras simuladas.

	Análise separada		**Análise conjunta**	
Parâmetro	*Média a posteriori*	*IC 95%*	*Média a posteriori*	*IC 95%*
	Submodelo longitudinal			*Submodelo longitudinal*
Ord. orig. (β_{11})	17.39	(17.12,17.64)	17.47	(17.21, 17.73)
tempo (β_{12})	1.81	(1.72,1.90)	1.71	(1.61, 1.80)
sexo (β_{13})	-0.63	$(-0.91, -0.30)$	-0.65	$(-0.94, -0.35)$
idade (β_{14})	-0.51	$(-0.94, -0.04)$	-0.58	$(-1.06, -0.15)$
IOPrev (β_{15})	-2.01	$(-2.34, -1.72)$	-2.12	$(-2.41, -1.81)$
σ^2	7.04	(6.87,7.22)	7.04	(6.86, 7.21)
σ^b_{11}	26.92	(25.74,28.25)	26.85	(25.64, 28.08)
σ^b_{22}	5.20	(4.84,5.60)	5.17	(4.77, 5.53)
cor $= \sigma^b_{12}/\sqrt{\sigma^b_{11}\sigma^b_{22}}$	-0.398	$(-0.44, -0.36)$	-0.36	$(-0.40, -0.33)$
	Submodelo de sobrevivência			*Submodelo de sobrevivência*
ord. orig. (β_{21})	-4.30	$(-4.54, -4.07)$	-5.03	$(-5.39, -4.67)$
sexo (β_{22})	0.33	(0.09, 0.58)	0.40	(0.14,0.65)
idade (β_{23})	0.62	(0.35,0.89)	0.76	(0.47,1.06)
IOPrev (β_{24})	0.87	(0.65,1.11)	1.02	(0.77,1.26)
γ_1	–	–	-0.23	$(-0.26, -0.2)$
γ_2	–	–	-0.38	$(-0.49, -0.26)$
σ^2_Q	–	–	0.014	(0.0001,0.038)

Tabela 2: Modelo separado (Modelo IV) e modelo conjunto (Modelo IX).

A Tabela 2 resume as estimativas *a posteriori* dos parâmetros de interesse e o respectivo intervalo de credibilidade 95% HPD (*Highest Posterior Density*), verificando-se efeito significativo de todas as covariáveis. Conclui-se ainda que i) os homens têm contagens de CD4 inferiores às das mulheres ($\beta_{13} < 0$),

ii) pacientes acima dos 50 anos e pacientes com alguma doença oportunista prévia à entrada no estudo têm valores de CD4 menores que os pacientes com menos de 50 anos e pacientes sem qualquer doença oportunista antes da entrada no estudo, respectivamente. No submodelo de sobrevivência é possível verificar que o risco de morrer é significativamente mais elevado nos homens do que nas mulheres, o que está de acordo com o facto de os homens terem valores de CD4 mais baixos. Existe um decréscimo no risco de morte no grupo abaixo dos 50 anos e um acréscimo de risco nos pacientes que já tiveram alguma doença oportunista. As estimativas *a posteriori* para γ_1 e γ_2 indicam que o nível inicial e a tendência temporal das contagens de CD4 estão negativamente associados com o risco de morte.
A Figura 1 mostra o gráfico do risco relativo, $\exp\{Q_k\}$, dos diferentes estados. Apesar da variação espacial exibida, os intervalos de credibilidade 95% HPD para este risco relativo não indicam uma significativa variação espacial. Do ponto de vista epidemiológico, isto pode mostrar que o combate à doença não difere entre os diversos estados Brasileiros.

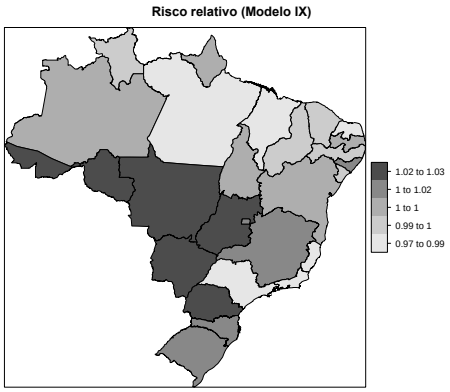

Figura 1: Mapa do Brasil onde são visíveis as heterogeneidades espaciais correlacionadas.

A análise conjunta bayesiana apresenta melhorias em relação ao tempo mediano de sobrevivência relativamente à separada. Para ilustrar esta situação consideremos dois pacientes censurados residentes no mesmo estado: (i) indivíduo número 61, masculino com 29 anos sem doenças oportunistas prévias e com um tempo de observação de 1624 dias; (ii) indivíduo número 62, feminino com 24 anos sem doenças oportunistas prévias e com um tempo de observação de 1390 dias. A Figura 2 mostra que o paciente 61 tem uma trajectória de CD4 mais favorável que a do paciente 62. Os resultados conjuntos diferem substancialmente dos resultados em separado, aumentando o tempo mediano de sobrevivência para o paciente 61, e diminuindo-o para o paciente 62. Além disso, revertem as evidências do modelo separado, no sentido em que o paciente com a "boa trajectória de CD4" está previsto sobreviver muito mais tempo do que o paciente com a "má trajectória".

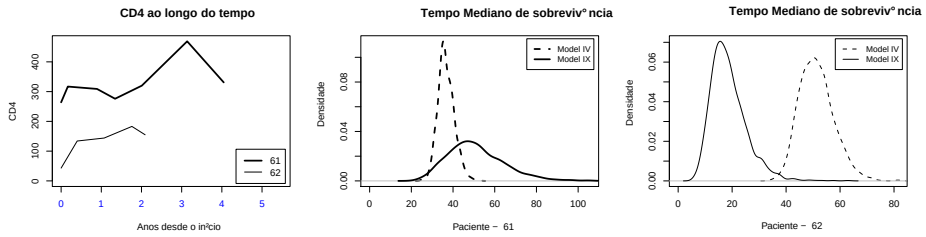

Figura 2: Trajectórias de CD4 dos indivíduos 61 e 62 (painel à esquerda); distribuição *a posteriori* do tempo mediano de sobrevivência para os dois pacientes (painel ao centro e painel à direita).

5 Avaliação do modelo

Para analisar o ajuste do modelo, no que toca à sobrevivência, analisámos as estimativas *a posteriori* dos resíduos Cox-Snell e Martingale. No que concerne à parte longitudinal foram considerados os resíduos marginais padronizados,

$$r_{ik}^{ym} = \hat{V}_{ik}^{-1/2}(\mathbf{y}_{ik} - \mathbf{X}_{1ik}\hat{\boldsymbol{\beta}}_1), \tag{8}$$

e os resíduos específicos padronizados,

$$r_{ik}^{ys}(t_{ikj}) = (y_{ik}(t_{ikj}) - \mathbf{x}_{1ik}^{\top}(t_{ikj})\hat{\boldsymbol{\beta}}_1 - \mathbf{z}_{1ik}^{\top}\hat{\boldsymbol{b}}_{ik}) \times \hat{\sigma}^{-1}. \tag{9}$$

Contudo, como notado por Rizopoulos et al. [5], o uso dos resíduos tradicionais nos modelos conjuntos não é viável, porque a sua distribuição de referência não está directamente disponível. Seguindo então uma ideia destes autores, combinámos num gráfico os resíduos observados correspondentes às observações, $\mathbf{y}_{ik}^{\text{obs}}$, com os resíduos de imputação múltipla, correspondentes às observações omissas, $\mathbf{y}_{ik}^{\text{omi}} = \{y_{ik}(t_{ikj}) : t_{ikj} \geq T_{ik}, j = 1, \ldots, m'_{ik}\}$, mas que foram imputadas com base na distribuição preditiva *a posteriori* da parte omissa do vector das respostas longitudinais, $\mathbf{y}_{ik}^{\text{omi}}$. O benefício que advém de usar valores simulados da resposta longitudinal após o abandono do indivíduo (por censura, morte ou outro), é que os resíduos vão herdar as propriedades do modelo para os dados completos e, portanto, podem ser usados directamente para produzir gráficos de diagnóstico sem ter de se considerar o abandono (não aleatório) dos pacientes. Caso os tempos de visita dos pacientes sejam aleatórios, como é o caso, o processo de visitas tem de ser tido em conta.

O gráfico dos resíduos específicos padronizados de cada sujeito na Figura 3 mostra uma tendência sistemática crescente do valor dos resíduos observados (linha cinzenta sólida), mas mostra também que este comportamento é "aliviado" quando consideramos os resíduos imputados (linha cinzenta tracejada). Portanto, a homocedasticidade dos erros $e_{ik}(t_{ikj})$ é uma condição que se verifica. No gráfico dos resíduos marginais padronizados observamos

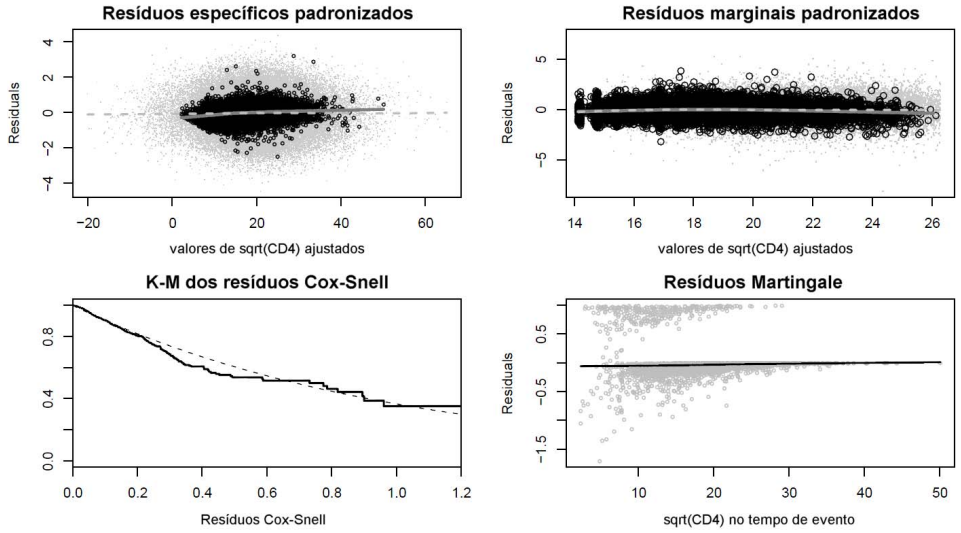

Figura 3: Painéis superiores - resíduos específicos padronizados e resíduos marginais padronizados (círculos pretos) aumentados com os resíduos imputados (pontos cinzentos). As linhas sólidas cinzentas representam um ajuste pelo "loess" baseado apenas nos resíduos observados e as linhas tracejadas sobrepostas representam um ajuste ponderado usando o "loess"; Painel inferior esquerdo - estimativas de Kaplan-Meier das estimativas *a posteriori* dos resíduos de Cox-Snell. A linha cinzenta tracejada é a distribuição exponencial unitária; Painel inferior direito - resíduos de Martingale *vs* valores ajustados de $\sqrt{\text{CD4}}$ calculados nos tempos de evento observados.

que a curva "loess" baseada apenas nos dados observados *versus* os valores ajustados de $\sqrt{\text{CD4}}$ mostra também uma ligeira tendência crescente (linha sólida), mas esse comportamento não está presente quando olhamos para os resíduos imputados (linha tracejada), indicando que depois de se ter em conta o abandono, o modelo ajustado conjunto parece ser um modelo plausível para este conjunto de dados. Os resíduos Martingale mostram que a forma funcional considerada para a relação entre a variável longitudinal e a função de risco é adequada, porque a linha obtida através do "loess" não mostra discrepâncias significativas relativamente a zero. O gráfico das estimativas Kaplan-Meyer para os resíduos de Cox-Snell não evidencia grandes desvios relativamente à distribuição exponencial unitária (linha tracejada), pois a maioria dos pontos estão perto da curva. Apesar de a parte central da curva apresentar um ligeiro desvio, este representa apenas cerca de 7% do total das observações.

Procedeu-se também a uma análise de sensibilidade, mas as estimativas *a posteriori* com base no modelo IX mostraram não ser sensíveis às diversas informações *a priori* testadas. Nomeadamente, investigou-se a influência da especificação da distribuição dos hiperparâmetros para a componente de precisão espacial, $\tau_Q = 1/\sigma_Q^2$, tendo sido admitida uma va-

riedade de distribuições *a priori* $\mathcal{G}(a,b)$. Em particular, o nosso delineamento experimental utilizou as seguintes combinações sugeridas em Silva *et al.* [8]: (a,b) = (0.5,0.0005), (0.001,0.001), (0.01,0.01), (0.1,0.1), (2,0.001), (0.2,0.0004) e (10,0.25).

6 Conclusões

O modelo conjunto seleccionado tem fragilidades espaciais, contudo, uma análise mais detalhada dos riscos relativos espaciais não mostrou haver diferenças significativas entre os 27 estados brasileiros. Apesar da complexidade deste modelo conjunto bayesiano, este é relativamente simples de implementar, existindo, claro, algumas limitações, por exemplo, a lenta convergência dos algoritmos MCMC que se traduz num enorme tempo computacional exigido, nomeadamente quando estão em causa efeitos aleatórios espaciais. O nosso modelo seleccionado (IX) levou cerca de 9 horas a ser processado num CPU com 4 núcleos. Uma abordagem recente, alternativa aos métodos MCMC, introduzida por Rue et al. [7] e conhecida por INLA (*Integrated Nested Laplace Approximations*), não se revelou ser mais rápida na nossa modelação conjunta, mas tal pode estar relacionado com o facto de os valores iniciais para os parâmetros dos modelos conjuntos utilizados no algoritmo MCMC serem o resultado da aplicação dos modelos em separado.

Agradecimentos

Os autores agradecem a Maria Medeiros e Cláudia Coeli pela base de dados, bem como ao revisor anónimo pelos seus comentários construtivos. Este trabalho foi parcialmente financiado pelos projectos PTDC/MAT/118335/2010 e Pest-OE/MAT/UI0006/2011.

Referências

[1] Guo, X., Carlin, P. (2004). Separate and joint modelling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58, 16-24.

[2] Henderson, R., Diggle, P., Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1, 465–480.

[3] Ibrahim, J.G., Chen, M.H., Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag.

[4] Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

[5] Rizopoulos, D., Verbeke, G., Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, 66, 20–29.

[6] Rizopoulos, D. (2012). Joint models for longitudinal and time-to-event data with applications in R. *Chapman and Hall/CRC*.

[7] Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society - B*, 71, 319–392.

[8] Silva, G.L., Dean, C.B., Niyonsenga, T., Vanasse, A. (2008). Hierarchical Bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in Medicine*, 27, 2381–2401.

[9] Souza-Jr, P.R.B., Szwarcwald, C.L., Castilho, E.A. (2007). Delay in introducing antiretroviral therapy in patients infected by HIV in Brazil, 2003-2006. *Clinics*, 62, 579–584.

[10] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* 64 (4). 583–639.

[11] Tsiatis, A.A., Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14, 809–834.