XXII CONGRESSO
SPE2015
SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

Estatística Bayesiana Computacional
-uma introdução

M. Antónia Amaral Turkman
Carlos Daniel Paulino

SPE
SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

Estatística Bayesiana Computacional

- uma introdução

M. Antónia Amaral Turkman
CEAUL & Faculdade de Ciências, Universidade de Lisboa
maturkman@fc.ul.pt

Carlos Daniel Paulino
CEAUL & Instituto Superior Técnico, Universidade de Lisboa
daniel.paulino@tecnico.ulisboa.pt

Edições SPE

FICHA TÉCNICA:

Livro: Estatística Bayesiana Computacional - uma introdução

Autores: M. Antónia Amaral Turkman
CEAUL & Faculdade de Ciências, Universidade de Lisboa

Carlos Daniel Paulino
CEAUL & Instituto Superior Técnico, Universidade de Lisboa

Editora: Sociedade Portuguesa de Estatística

Impressão: Instituto Nacional de Estatística

Tiragem: 300 exemplares

ISBN: 978-972-8890-37-7

Depósito Legal: 395649/15

Prefácio

Em 1975, Dennis Lindley escreveu um artigo em *Advances in Applied Probability* intitulado *The Future of Statistics – a Bayesian 21st Century*, prevendo o predomínio para o século XXI da abordagem bayesiana como metodologia inferencial em Estatística. Hoje pode certamente dizer-se que Dennis Lindley acertou na sua previsão, embora não exatamente pelas razões por ele preconizadas, devido ao grande avanço registado durante a última década do século XX da denominada Estatística Bayesiana Computacional. É certo que a "solução bayesiana" para os problemas de inferência é altamente atrativa, particularmente no que diz respeito à interpretabilidade das inferências resultantes. Contudo, na prática, a obtenção dessa solução passa dominantemente pela necessidade de calcular integrais, na maioria dos casos multidimensionais, não sendo portanto fácil, se não impossível, executá-la sem recurso ao computador. O desenvolvimento de métodos computacionais, mais ou menos sofisticados, veio mudar por completo o panorama. Hoje os métodos bayesianos são usados para resolver problemas em praticamente todas as áreas científicas, particularmente quando os processos a modelar são extremamente complexos. Contudo, a aplicação da metodologia bayesiana não pode ser feita cegamente. Apesar de existir atualmente muito software de análise bayesiana, é absolutamente necessário que se perceba o que se está a produzir, como e porquê.

O objetivo deste texto, associado ao minicurso lecionado no XXII Congresso da Sociedade Portuguesa de Estatística, é precisamente o de apresentar as ideias fundamentais que estão subjacentes à formulação e análise dos modelos bayesianos, dando particular relevo a esquemas e meios computacionais que as permitem realizar.

Começa-se por apresentar no Capítulo 1 uma breve resenha sobre os fundamentos da inferência bayesiana com referência às principais diferenças entre os paradigmas clássico e bayesiano. Como uma das pedras basilares da inferência bayesiana, a quantificação da informação *a priori*, infelizmente tantas vezes ignorada nas aplicações, é uma questão que será também abordada no Capítulo 2 nos seus aspetos essenciais. No Capítulo 3 exemplos analiticamente tratáveis são usados para ilustrar a solução bayesiana a problemas de inferência estatística. A "grande ideia" por trás do desenvolvimento da Estatística Bayesiana Computacional é o reconhecimento de que as inferências bayesianas podem ser feitas por recurso a amostras simuladas da distribuição *a posteriori*. Os métodos clássicos de Monte Carlo são apresentados no Capítulo 4, como um primeiro recurso para resolver problemas computacionais com que de imediato o investigador se depara, mesmo em simples contextos uniparamétricos. A avaliação de modelos é uma questão muito importante e que tem a sua filosofia própria no contexto bayesiano. Os métodos mais usados para a crítica, seleção e comparação de modelos são resumidamente abordados no Capítulo 5.

Situações mais complexas do que as abordadas no Capítulo 4 exigem o recurso a métodos de simulação mais sofisticados, nomeadamente a métodos de Monte Carlo via cadeias de Markov (MCMC). Estes são apresentados no Capítulo 6 de um modo tão simples quanto possível. A possibilidade de recurso a estes métodos para amostrar da distribuição *a posteriori*, a par do desenvolvimento do software BUGS, permitiu a aplicação da metodologia bayesiana a uma grande variedade de problemas e a sua expansão a outras áreas científicas. Os avanços verificados nos instrumentos e tecnologias em geral têm vindo a mudar o paradigma da Estatística, havendo hoje a necessidade de lidar com quantidades massivas de dados ("Big Data Era"), muitas vezes de natureza espacial e temporal. Como consequência, simular da distribuição *a posteriori* em problemas com dados de natureza complexa e de grande dimensão passou a ser um novo desafio, o qual veio acompanhado de novos e melhores métodos computacionais e do desenvolvimento de software mais adequado para ultrapassar as limitações do BUGS e seus sucessores, WinBUGS e OpenBUGS. Neste livro são também abordados no Capítulo 8 outros pacotes estatísticos que implementam métodos MCMC e suas variantes, como sejam JAGS, STAN e BayesX. Outra alternativa à simulação é a utilização de métodos de aproximação da distribuição *a posteriori*, tema que se aborda sucintamente no Capítulo 7. Nesse quadro descreve-se neste capítulo, de um modo genérico, a

abordagem por aproximações de Laplace encaixadas e integradas (INLA), a qual permite ganhos quer em tempo computacional (por vezes enormes), quer na precisão das inferências efetuadas. Embora o tipo de problemas que podem ser manejados com esta metodologia seja vasto, é bem mais limitado do que aqueles tratados por métodos de simulação estocástica. Aborda-se ainda no Capítulo 8 reservado ao software bayesiano, aspetos relevantes da abordagem INLA com uma breve referência ao pacote do R que a implementa (R-INLA).

Com certeza que para a execução deste texto nos baseámos no livro de Estatística Bayesiana de 2003, editado pela Fundação Calouste Gulbenkian, de que somos autores em parceria com Bento Murteira. Dado que este livro se encontra esgotado há muito tempo, também aproveitámos bastante do atual produto do trabalho preparatório da sua 2ª edição, bem como do material por nós publicado na edição de Outono de 2013 do boletim da Sociedade Portuguesa de Estatística (SPE).

Este texto não teria visto a luz do dia, no presente formato, sem a preciosa e incansável ajuda do nosso muito amigo e colega Giovani Silva. Para ele vão os nossos mais sinceros agradecimentos. Desejamos também manifestar a nossa gratidão à Sociedade Portuguesa de Estatística por ter proposto o amplo tema de Estatística Bayesiana e nos ter dado a oportunidade de o lecionar no minicurso do seu XXII Congresso. Agradecemos ainda o apoio institucional da Universidade de Lisboa através do Centro de Estatística e Aplicações (PEst-OE/MAT/UI0006/2014, UID/MAT/00006/2013), do Departamento de Estatística e Investigação Operacional da Faculdade de Ciências e do Departamento de Matemática do Instituto Superior Técnico. Finalmente, queremos deixar registado que a investigação que nos permitiu atingir a nossa plataforma de conhecimento na área de Estatística Bayesiana foi sempre parcialmente subsidiada pela Fundação para a Ciência e Tecnologia, através de diversos projetos ao longo de muitos anos.

Por fim, queremos dedicar este livro ao Professor Bento Murteira a quem muito se deve a divulgação em Portugal das ideias fundamentais da Estatística Bayesiana – o próprio Capítulo 1 deste livro reflete bem o sabor de alguns dos seus úteis manuscritos.

Lisboa, 20 de Setembro de 2015

M. Antónia Amaral Turkman

Carlos Daniel Paulino

Índice

Capítulo 1

Fundamentos da Inferência Bayesiana

1.1 O problema fundamental da Estatística

Antes de abordar os alicerces da inferência bayesiana parece conveniente fazer referência ao problema fundamental da Estatística. Para O'Hagan: "*The fundamental problem towards which the study of statistics is addressed is that of inference. Some data are observed and we wish to make statements, inferences, about one or more unknown features of the physical system which gave rise to these data*". Ao aprofundar o estudo dos fundamentos da Estatística depara-se com grande número de correntes ou escolas. Sem falar nos chamados clássicos o desfile é extenso: bayesianos[1] (objetivos, subjetivos[2], ...), estruturalistas, fiducialistas, verosimilhancistas, ...

A diversidade não é inesperada! As informações dos dados sobre parâmetros ou modelos enquadram-se, em geral, na indução que é um dos problemas mais controversos da filosofia. Cada escola tem princípios e procedimentos próprios cuja análise conduz aos fundamentos da inferência estatística con-

[1] Até há meios-bayesianos que consideram relevante utilizar a informação *a priori* mas entendem que para a combinar com os dados a linguagem das probabilidades é inadequada e deve ser substituída por uma que atenda às relações causais.

[2] A que se dedica essencialmente o presente volume.

forme Berger descreve: "*Statistics needs a: 'foundation', by which I mean a framework of analysis within which any statistical investigation can theoretically be planned, performed, and meaningfully evaluated. The words 'any' and 'theoretically' are key, in that the framework should apply to any situation but may only theoretically be implementable. Practical difficulties or time limitations may prevent complete (or even partial) utilisation of such framework, but the direction in which 'truth' could be found would at least be known*". Os fundamentos da inferência bayesiana são melhor compreendidos quando introduzidos em confronto com os da principal "concorrente", a inferência clássica.

1.2 O paradigma clássico

A inferência clássica procura determinar que generalizações sobre a população podem fazer-se a partir da amostra. Representando os dados estatísticos por x [ou $x = (x_1, x_2, \ldots, x_n)$ onde n é a dimensão da amostra], o conjunto $\mathcal{X}$ de amostras possíveis x designa-se por espaço amostral (ou espaço-amostra). Correntemente, $\mathcal{X} \subseteq \Re$ ou $\mathcal{X} \subseteq \Re^n$. Fundamental para inferência clássica é reconhecer a variabilidade que se verifica de amostra para amostra tendo em mente que os dados observados formam apenas um dos muitos – possivelmente infinitos – conjuntos que poderiam ter sido obtidos. A interpretação dos dados depende não apenas da particular amostra observada mas também das hipóteses adotadas acerca das possíveis amostras alternativas. Em consequência os dados consideram-se como observação de uma variável ou vetor aleatório X com função de distribuição F_0, que não é, como é óbvio, perfeitamente conhecida. No entanto, existe normalmente algum conhecimento inicial (considerações teóricas, evidência experimental, etc.) sobre a natureza do fenómeno aleatório em estudo que leva a propor ou conjeturar uma família de distribuições $\mathcal{F}$ a que pertence F_0 e que se designa por modelo estatístico para X. A proposta de um modelo é conhecida por *especificação* e é uma fase essencial no estabelecimento de inferências.

Admitindo que X é uma variável ou vetor aleatório contínuo, é prática corrente representar as distribuições de $\mathcal{F}$ pelas respetivas funções densidade de probabilidade. Se estas forem rotuladas por um parâmetro θ com domínio no espaço paramétrico (ou espaço-parâmetro) Θ, o modelo estatístico pode escrever-se $\mathcal{F} = \{f(x|\theta),\, x \in \mathcal{X} : \theta \in \Theta\}$. Em muitos casos, as n variáveis aleatórias $(X_1, X_2, \ldots, X_n)$ são supostas independentes condicionalmente em θ e

o modelo estatístico pode representar-se em termos das densidades marginais das variáveis $X_i, i = 1, 2, \ldots, n$,

$$\mathcal{F} = \left\{ f(x|\theta) = \prod_{i=1}^{n} f_i(x_i|\theta) : \theta \in \Theta \right\}, x \in \mathcal{X},$$

onde $f_i(\cdot|\theta) = f(\cdot|\theta), i = 1, 2, \ldots, n$, se adicionalmente as variáveis, X_i forem identicamente distribuídas, situação frequentemente denominada de amostragem casual ou aleatória.

A escolha da família $\mathcal{F}$ resulta de uma síntese entre vários fatores, designadamente a prévia evidência experimental decorrente de tratamento de fenómenos análogos, as considerações teóricas sobre os objetivos do estudo, a natureza dos fenómenos em questão e das técnicas experimentais aplicadas e ainda os requisitos de parcimónia e interpretabilidade.

Ultrapassada a tarefa de modelação e parametrização, a inferência clássica contém inúmeros procedimentos para extrair da amostra conclusões sobre as características do modelo representativo da população e procura responder a questões como estas: (a) Os dados x sustentam ou são compatíveis com a família $\mathcal{F}$? (b) Supondo que a especificação está correta e que os dados emanam de uma das distribuições da família $\mathcal{F}$, que conclusões podem tirar-se sobre o valor particular do parâmetro θ_0 que indexa a função de distribuição F_0 que descreve "apropriadamente" os fenómenos investigados?

Os procedimentos clássicos – também designados frequencistas – são avaliados à luz do princípio da *amostragem repetida* através da consideração do seu comportamento num número indefinido de hipotéticas repetições efetuadas nas mesmas condições. Uma das faces do princípio reside precisamente na utilização de frequências como medidas de incerteza, *i.e.* na interpretação frequencista de probabilidade (*vide* Paulino, Amaral Turkman e Murteira, 2003, Sec. 1.2, para uma resumida descrição desta e de outras interpretações do conceito de probabilidade).

Considerando o caso paramétrico, na resposta à questão (b) acima há que considerar em primeiro lugar a estimação pontual que consiste, *grosso modo*, no seguinte: dada a amostra $X = (X_1, X_2, \ldots, X_n)$, como "adivinhar", conjeturar ou aproximar o verdadeiro valor do parâmetro θ através do emprego de um estimador $T(X_1, X_2, \ldots, X_n)$, que os clássicos desejam que seja dotado de propriedades interessantes tais como não enviesamento, consistência, suficiência, eficiência, etc.

Por exemplo, com $\mathcal{X} \equiv I\!R^n$, o estimador $T(X_1, X_2, \ldots, X_n)$ baseado numa amostra casual diz-se centrado ou não enviesado quando verifica,

$$E\{T \mid \theta\} = \int_{I\!R^n} T(x_1, x_2, \ldots, x_n) \prod_{i=1}^{n} f(x_i|\theta) dx_1 dx_2 \ldots dx_n = \theta, \ \forall \theta \in \Theta.$$

Trata-se, claramente, de uma propriedade estabelecida em conformidade com o princípio da amostragem repetida, porquanto integrar sobre o espaço amostral (no caso presente $I\!R^n$), i.e., percorrer todo esse espaço, só é possível quando se concebem ou imaginam infinitas repetições do processo de amostragem ou observação das n variáveis aleatórias $(X_1, X_2, \ldots, X_n)$. O mesmo acontece quando se consideram as outras propriedades de avaliação de estimadores no paradigma em questão. Por outras palavras, no princípio da amostragem repetida encontra-se nítida referência ao que se passa em todo o espaço da amostra.

A estimação paramétrica assume muitas vezes a forma de intervalos de confiança. Em vez de propor um único valor para θ indica-se um intervalo cujos limites dependem da amostra,

$$(T^*(X_1, X_2, \ldots, X_n), T^{**}(X_1, X_2, \ldots, X_n))$$

e enquadram o verdadeiro valor do parâmetro com um certa probabilidade que interessa ser elevada (tipicamente, o chamado grau de confiança)

$$P\{T^*(X_1, X_2, \ldots, X_n) < \theta < T^{**}(X_1, X_2, \ldots, X_n) \mid \theta\} = 1 - \alpha, \ 0 < \alpha < 1.$$

Esta expressão traduz pré-experimentalmente a probabilidade de cobertura do valor desconhecido de θ pelo intervalo aleatório (T^*, T^{**}) cujos limites inferior e superior são funções de $(X_1, X_2, \ldots, X_n)$ e, portanto, variáveis aleatórias. No entanto, quando se recolhe uma amostra concreta (i.e., pós-experimentalmente) e se obtêm n números reais, $(x_1, x_2, \ldots, x_n)$, fica determinado um intervalo do eixo real (os limites inferior e superior são agora números reais),

$$(T^*(x_1, x_2, \ldots, x_n), T^{**}(x_1, x_2, \ldots, x_n)),$$

mas a seguinte probabilidade,

$$P\{^*T(x_1, x_2, \ldots, x_n) < \theta < T^{**}(x_1, x_2, \ldots, x_n) \mid \theta\} = 1 - \alpha, \ 0 < \alpha < 1,$$

deixa de fazer sentido. De facto, uma vez que θ tem um valor desconhecido, mas fixo, essa probabilidade só pode ter o valor 1 ou 0, conforme o verdadeiro valor de θ pertence ou não ao intervalo numérico $(T^*(x_1, x_2, \ldots, x_n), T^{**}(x_1, x_2, \ldots, x_n))$. Mas, evidentemente, como o parâmetro θ é desconhecido, o investigador não sabe em que situação está. No entanto, sendo um estatístico clássico, aceita a interpretação frequencista de probabilidade e invoca o princípio da amostragem repetida do seguinte modo: se conceber a repetição do processo de amostragem (cada amostra envolve n observações) um grande número de vezes, em $(1-\alpha)\times 100\%$ dos casos o intervalo numérico inclui o valor de θ.

Outra modalidade de inferência estatística clássica é o teste de hipóteses paramétricas. No decurso da investigação científica há muitas vezes lugar, no quadro de determinada teoria, à formulação de hipóteses sobre o valor de um (ou de vários) parâmetro(s), por exemplo, simbolicamente,

$$H_0 : \theta = \theta_0.$$

A questão fundamental que se levanta é a seguinte: os dados $(x_1, x_2, \ldots, x_n)$ sustentam ou não a hipótese proposta tradicionalmente designada por hipótese de nulidade? Também aqui a solução clássica continua a basear-se no princípio da amostragem repetida quando segue a teoria de Neyman-Pearson e procura definir uma região de rejeição W – região crítica – definida no espaço amostral, $W \subset \mathcal{X}$, tal que,

$$(X_1, X_2, \ldots, X_n) \in W \;\Rightarrow\; \text{rejeição de } H_0,$$

$$(X_1, X_2, \ldots, X_n) \in \overline{W} \;\Rightarrow\; \text{aceitação de } H_0,$$

com $\overline{W} = \mathcal{X} - W$, e pretende controlar a probabilidade de um erro de 1ª espécie,

$$P\{(X_1, X_2, \ldots, X_n) \in W | H_0 \text{ verdadeira}\}$$

e minimizar a probabilidade de um erro de 2ª espécie,

$$P\{(X_1, X_2, \ldots, X_n) \in \overline{W} | H_0 \text{ falsa}\}.$$

O que quer dizer que determinada região crítica tem associada uma probabilidade de um erro de 1ª espécie igual, por exemplo, a 0.05? Se uma particular observação cai na região crítica e a hipótese é consequentemente rejeitada,

o investigador não sabe de facto se está a rejeitar uma hipótese falsa ou uma hipótese verdadeira. No entanto, sendo um clássico, está convicto de que se conceber a repetição do procedimento um número indeterminado de vezes e se a hipótese for verdadeira, só em 5% dos casos é que a observação cai na região crítica e é por isso rejeitada. O que quer dizer que determinada região crítica tem associada uma probabilidade de cometer um erro de 2ª espécie igual a 0.10? De modo análogo, se a particular observação não pertence à região crítica e, consequentemente, a hipótese é aceite, o investigador não sabe se está a aceitar uma hipótese verdadeira ou falsa. Sendo um clássico, afirma-se convencido que se repetir um número indeterminado de vezes o procedimento e se a hipótese for falsa, só em 10% dos casos é que a observação não pertence à região crítica e é com tal aceite.

A forma como a inferência clássica perspetiva os problemas de estimação e de ensaio de hipóteses supõe-se conhecida, pelo menos nos aspetos mais elementares, e, por isso, a presente análise não se alonga mais.

1.3 O paradigma bayesiano

Para Lindley a substituição do paradigma clássico pelo paradigma bayesiano representa uma verdadeira revolução científica no sentido de Kuhn[3]. A semente para a abordagem bayesiana a problemas de inferência foi lançada por Richard Price quando em 1763 publicou a obra póstuma do Rev. Thomas Bayes intitulada "*An Essay Towards Solving a Problem in the Doctrine of Chances*". A interpretação da probabilidade como grau de credibilidade – fundamental na filosofia bayesiana – tem uma longa história. Se parece ter sido J. Bernoulli, em 1713, na sua obra *Ars Conjectandi*, um dos primeiros autores a definir probabilidade como o grau de confiança na veracidade de uma dada proposição, foi De Morgan, na *Formal Logic*, em 1847, que afirmou: (1) a probabilidade identifica-se com um grau de credibilidade; (2) os graus de credibilidade podem medir-se; (3) os graus de credibilidade podem identificar-se com um certo complexo de sentimentos. A ideia de coerência de um sistema de graus de credibilidade parece dever-se a Ramsey para quem a atitude de um indivíduo ao apostar na veracidade de uma dada proposição está associada com o grau de credibilidade que lhe atribui. Se declara

[3]Kuhn, T.S. The Structure of Scientific Revolutions. Chicago: University of Chicago Press, 1962.

que as vantagens ou possibilidades (chances) – a favor da veracidade contra a não veracidade – são $r : s$, então o grau de credibilidade da proposição é, para o indivíduo, $r/(r + s)$. Para Ramsey nenhum conjunto de apostas num grupo de proposições é admitido a um indivíduo coerente se conduzir a um prejuízo certo. O maior expoente do conceito de probabilidade personalista é, contudo, de Finetti. Nas ideias bayesianas e na sua aplicação à estatística tem de citar-se Harold Jeffreys que, reagindo à predominante posição clássica em meados do século, embora solitário e desapoiado, conseguiu ressuscitar o bayesianismo, dar-lhe *status* lógico e avançar com soluções de problemas estatísticos que naquele tempo persistiam. A partir daí a lista de bayesianos foi engrossando sucessivamente e, na impossibilidade de citar todos, merecem realce os nomes de Good, Savage e Lindley.

O Teorema de Bayes, bem conhecido, é uma proposição sobre probabilidades condicionadas indiscutível decorrente do cálculo de probabilidades ou da axiomática de Kolmogorov. O aspecto controverso é a sua aplicação a problemas de inferência estatística. Ocupa, como é óbvio, lugar fulcral na inferência bayesiana que tem relativamente à inferência clássica uma divergência fundamental. No modelo clássico o parâmetro $\theta, \theta \in \Theta$, é um escalar ou vetor desconhecido, mas fixo, *i.e.*, igual ao valor particular que indexa a distribuição da família $\mathcal{F}$ que descreve "apropriadamente" o processo ou sistema físico que gera as observações. No modelo bayesiano o parâmetro $\theta, \theta \in \Theta$, é tomado como um escalar ou vetor aleatório não observável. Segundo os bayesianos o que é desconhecido – no caso em questão, o parâmetro θ – é incerto e toda a incerteza deve ser quantificada em termos de probabilidade. Correlativamente, os bayesianos defendem que a informação inicial ou *a priori* – anterior ou externa em relação à experiência mas demasiado importante para ser ignorada, pois pode até suceder que os investigadores sejam peritos na matéria – deve traduzir-se por uma distribuição de probabilidade para θ, seja $h(\theta)$, designada distribuição *a priori*. A determinação e a interpretação da distribuição *a priori* são dos pontos mais controversos da teoria bayesiana.

A família $\mathcal{F}$ também faz parte do modelo bayesiano; quer dizer, a componente amostral ou experimental é comum aos modelos clássico e bayesiano, embora para este os elementos $f(x|\theta)$ de $\mathcal{F}$ em geral são supostos ter, tal como $h(\theta)$, uma interpretação subjetiva.

A discussão das distribuições *a priori* ilustra alguns aspetos do confronto entre bayesianos e clássicos. Para os primeiros, Berger por exemplo, a es-

pecificação – escolha subjetiva da família $\mathcal{F}$ – traduz muitas vezes um uso mais drástico da informação *a priori* do que o emprego de uma distribuição *a priori*. E acrescentam: na sua modelação os clássicos atendem à informação *a priori* quando muito informalmente, atitude que consideram algo limitada porquanto, para eles, a informação inicial ou *a priori* detida por um dado investigador deve traduzir-se formalmente por uma distribuição de probabilidade para a variável aleatória θ. Para os segundos, por exemplo Lehmann, há uma importante diferença entre a modelação de $\mathcal{F}$ e a modelação de $h(\theta)$ pois no primeiro caso dispõe-se de um conjunto de observações – $x = (x_1, x_2, \ldots, x_n)$ – geradas por um membro de $\mathcal{F}$ que pode empregar-se para testar a forma da distribuição.

Para entender o ponto de vista bayesiano repare-se que um clássico *em todos* os problemas que envolvem uma variável X com distribuição binomial recorre sempre ao modelo bernoulliano em que o parâmetro θ representa a probabilidade de um "sucesso". Para os bayesianos cada problema *é único* e tem um contexto real próprio onde θ é uma quantidade significativa acerca da qual existem, em geral, graus de conhecimento que variam de problema para problema e de investigador para investigador. Assim, a distribuição de probabilidade que capta essa variabilidade é baseada na informação *a priori* e é especifica de um dado problema e de um dado investigador. De facto, sublinham, a informação *a priori* inclui juízos ou experiências individuais da mais diversa índole, decorrentes em geral de situações não repetitivas, formalizáveis apenas em termos personalistas. No entanto, advertem, esta formalização exige que o investigador satisfaça condições de coerência ou de consistência que permitam o recurso às regras de cálculo de probabilidades estabelecidas. Assim, diferentes investigadores possuem em geral diferentes distribuições *a priori* para o mesmo parâmetro sem deixarem necessariamente de ser coerentes.

Suponha-se que se observa $X = x$; dado um qualquer $f(x|\theta) \in \mathcal{F}$ e a distribuição *a priori* $h(\theta)$, o Teorema de Bayes conduz à relação[4],

$$h(\theta|x) = \frac{f(x|\theta)h(\theta)}{\int_\Theta f(x|\theta)h(\theta)d\theta}, \ \theta \in \Theta, \tag{1.1}$$

onde $h(\theta|x)$ é a distribuição *a posteriori* de θ depois de conhecido $X = x$. Assim, a atitude inicial do investigador, caraterizada por $h(\theta)$, é modificada

[4]Facilmente adaptável se x for um vetor ou se o espaço do parâmetro for discreto.

pela informação recolhida passando a traduzir-se por $h(\theta|x)$. O denominador de (1.1), que se representa por $f(x)$, é a distribuição marginal (ou preditiva *a priori* de X; diz respeito à observação de X qualquer que seja θ.

O conceito de função de verosimilhança estuda-se no quadro da inferência clássica mas não é menos importante no quadro bayesiano. Na respectiva definição convém manter a distinção entre os casos discreto e contínuo [veja-se Kempthorne e Folks (1971)] mas chega-se em ambos os casos à função de θ,

$$\begin{array}{l} L(\theta|x) \;=\; kf(x|\theta), \quad \theta \in \Theta \quad \text{ou} \\ L(\theta|x_1,\ldots,x_n) \;=\; k\prod_i f(x_i|\theta), \quad \theta \in \Theta, \end{array} \tag{1.2}$$

que para cada $\theta \in \Theta$ exprime a verosimilhança ou plausibilidade que lhe é atribuída quando se observa $X = x$ ou $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. O símbolo k representa uma função que não depende de θ. A função de verosimilhança – não é uma probabilidade pois, por exemplo, não faz sentido adicionar verosimilhanças – tem importante papel na fórmula de Bayes pois representa o meio através do qual os dados, x, transformam o conhecimento *a priori* sobre θ; quer dizer, a verosimilhança pode interpretar-se como expressão da informação sobre θ fornecida pelos dados x.

Em resumo, para os bayesianos a distribuição *a posteriori* incorpora, por via do Teorema de Bayes, toda a informação disponível sobre o parâmetro,

informação inicial + informação da experiência ou da amostra.

Daqui decorre que todos os procedimentos de inferência bayesiana são baseados em $h(\theta|x)$ [ou $h(\theta|x_1, x_2, \ldots, x_n)$].

No caso em que θ representa um vector de parâmetros, *e.g.*, $\theta = (\gamma, \phi) \in \Gamma \times \Phi$, pode acontecer que o interesse inferencial se restrinja apenas a uma parte de θ, digamos γ. Nessa eventualidade, a eliminação do parâmetro perturbador ϕ no paradigma bayesiano obedece sempre à mesma via, contrariamente ao que se verifica no paradigma clássico, traduzida na marginalização da distribuição *a posteriori* conjunta, *i.e.*, na determinação de

$$h(\gamma|x) = \int_\Phi h(\gamma, \phi|x)d\phi = \int_\Phi h(\gamma|\phi, x)h(\phi|x)d\phi. \tag{1.3}$$

A possível dificuldade de integração analítica desaparece sempre que γ e ϕ são aprioristicamente independentes e a função de verosimilhança se fatoriza em $L(\theta|x) = L_1(\gamma|x) \times L_2(\phi|x)$, resultando que $h(\gamma|x) \propto h(\gamma)L_1(\gamma|x)$.

1.4 Inferência bayesiana

Nos procedimentos bayesianos podem distinguir-se dois objetivos: **I** – Traçar inferências sobre o parâmetro não observável θ; **II** – Realizar inferências sobre variáveis ainda não observadas (predição).

1.4.1 Inferências paramétricas

Na ótica da inferência paramétrica há uma certa coincidência – pelo menos superficial – entre os objetivos dos clássicos e dos bayesianos mas na sua implementação as duas correntes entram em choque. Por um lado, as inferências clássicas são baseadas em probabilidades associadas com as diferentes amostras, x, que poderiam ocorrer para algum valor fixo, mas desconhecido, do parâmetro θ. Isto é, as inferências têm por base as distribuições amostrais que "ponderam", probabilisticamente, os valores que a variável X ou a estatística $T(X)$ podem assumir percorrendo o espaço amostral. Por outro lado, as inferências bayesianas são baseadas em probabilidades subjetivas ou credibilidades *a posteriori* associadas com diferentes valores do parâmetro θ e condicionadas pelo particular valor de x observado. O ponto x está fixo e determinado e é a variação de θ que é considerada.

Por exemplo, os bayesianos uma vez observado x e considerando a hipótese de ser $\{\theta \leq 0.5\}$, respondem à questão de forma significativa e direta calculando $P(\theta \leq 0.5|x)$ a partir de $h(\theta|x)$, *i.e.*, sem sair do cálculo de probabilidades. Em contraste, os clássicos não respondem diretamente à questão e ao afirmarem, por exemplo, que a hipótese $H_0 : \theta \leq 0.5$ é rejeitada ao nível de 5% não querem afirmar que a sua probabilidade é inferior a 0.05, mas que se a hipótese H_0 for verdadeira (*i.e.*, se de facto $\theta \leq 0.5$), então a probabilidade de X pertencer a uma dada região crítica W a determinar é tal que $P(X \in W|\theta \leq 0.5) < 0.05$, e que se de facto $x \in W$, então tal hipótese deve ser rejeitada.

No dizer crítico de O'Hagan, enquanto os bayesianos podem emitir enunciados probabilísticos sobre os parâmetros, que consideram como variáveis aleatórias, isso não é possível com os clássicos cujas probabilidades dizem respeito aos dados e não ao parâmetro, embora depois reformuladas para que aparentemente digam respeito ao parâmetro. Esta questão tem correspondência na diferente atitude em relação ao espaço da amostra. Para os clássicos o conceito é fundamental pois a amostragem repetida consiste em "percorrer"

tal espaço. Os bayesianos começam por criticar a ideia de que é pacífico fazer repetições com n fixo; depois defendem que só interessa o resultado obtido, x, e não o conjunto ou espaço a que pertence x que pode ser absolutamente arbitrário e que contém, além de x, observações que poderiam ter sido obtidas mas que não o foram[5].

Nos problemas de estimação os clássicos consideram diferentes alternativas ou funções dos dados – estimadores – cujas propriedades amostrais investigam sob diversas óticas (consistência, não enviesamento, etc.). Para os bayesianos há apenas *um* estimador que é precisamente a distribuição *a posteriori* $h(\theta|x)$. Pode, é claro, descrever-se esta distribuição através da moda, média, mediana ou da variância, mas isso nada tem a ver com o problema que enfrentam os clássicos quando pretendem determinar o chamado *estimador ótimo*. Para os bayesianos tal problema só existe no quadro da Teoria da Decisão, campo em que os bayesianos têm nítida vantagem sobre os clássicos. Em consonância, Savage sustenta que nas últimas décadas o problema central em face da incerteza deixou de ser *que inferências podem realizar-se* e passou a ser *que fazer ou que decisão tomar*. Como a decisão individual tem sido considerada ultrapassada por alguns filósofos, fala-se ultimamente no reforço do bayesianismo através da decisão em grupo.

Aos intervalos de confiança os bayesianos contrapõem os intervalos (ou regiões) de credibilidade. Observado x e determinada a distribuição *a posteriori*, um intervalo de credibilidade para o parâmetro θ (suposto aqui um escalar) é formado por um par de valores de Θ, sejam $[\underline{\theta}(x), \bar{\theta}(x)]$, ou mais simplesmente, $(\underline{\theta}, \bar{\theta})$, tais que,

$$P(\underline{\theta} < \theta < \bar{\theta}|x) = \int_{\underline{\theta}}^{\bar{\theta}} h(\theta|x)d\theta = 1 - \alpha, \tag{1.4}$$

onde $1-\alpha$ (em geral , 0.90, 0.95 ou 0.99) é o nível de credibilidade desejado. Se $\Theta = (-\infty, +\infty)$ uma forma expedita de construir um intervalo de credibilidade (dito então central) é considerar na distribuição *a posteriori* abas de igual credibilidade verificando

$$\int_{-\infty}^{\underline{\theta}} h(\theta|x)d\theta = \int_{\bar{\theta}}^{+\infty} h(\theta|x)d\theta = \frac{\alpha}{2}. \tag{1.5}$$

[5] A relutância em relação ao espaço amostral traduz o mesmo sentimento acerca das regras de paragem na experimentação, tema que Mayo e Kruse (2001), recordando Armitage, dizem poder levantar aos bayesianos alguns problemas.

A definição (1.4) possui um inconveniente: o intervalo $(\underline{\theta}, \overline{\theta})$ não é único, podendo suceder que valores de θ contidos nesse intervalo tenham menor credibilidade que valores de θ não incluídos no mesmo intervalo. Assim, para proceder à escolha de um certo intervalo ao mesmo tempo que se minimiza a respetiva amplitude, os bayesianos preferem trabalhar com intervalos de credibilidade HPD (*highest posteriori density*) $(\theta', \theta'') = \{\theta : h(\theta|x_1, x_2, \ldots, x_n) \geq k(\alpha)\}$, onde $k(\alpha)$ é o maior número real tal que $P(\theta' < \theta < \theta'') = 1 - \alpha$.

Os intervalos de credibilidade têm uma interpretação direta em termos de probabilidade. O mesmo não se passa com os intervalos de confiança em que se parte de uma probabilidade que não diz respeito a θ, mas sim aos dados, mais precisamente a um intervalo aleatório definido a partir da amostra genérica e que depois de observar uma amostra concreta se reconverte na *confiança* de cobertura do valor desconhecido de θ pelo intervalo numérico resultante, o qual não pode interpretar-se em geral como uma probabilidade ou credibilidade referente a θ. Além de outros aspetos críticos que a teoria dos intervalos (ou regiões) de confiança suscita, conhecem-se os irónicos comentários de Lindley (1990) quando diz conhecer várias axiomáticas da teoria da probabilidade – por exemplo, as devidas a Savage, de Finetti ou Kolmogorov – mas não conhece nenhuma axiomática da *confiança*.

Os bayesianos quando, por exemplo, pretendem ensaiar uma hipótese composta $H_0 : \theta \in \Theta_0$ *versus* uma alternativa também composta $H_1 : \theta \in \Theta_1$, com $\Theta_0 \cap \Theta_1 = \varnothing, \Theta_0 \cup \Theta_1 = \Theta$, chegam a expressões em termos de probabilidades sobre Θ. Se o investigador tem uma distribuição, $h(\theta)$, $\theta \in \Theta$, traduzindo a credibilidade inicial que atribui aos diferentes valores do parâmetro, pode determinar as probabilidades *a priori* das hipóteses em confronto,

$$P(\Theta_0) = \int_{\Theta_0} h(\theta) d\theta, \qquad P(\Theta_1) = \int_{\Theta_1} h(\theta) d\theta \ .$$

Ao quociente $P(\Theta_0)/P(\Theta_1)$ costuma chamar-se rácio das vantagens ou razão das chances (*odds*) *a priori* de H_0 sobre H_1. Depois de realizar a experiência, que se supõe consistir na observação de x, e de determinar $h(\theta|x)$, o estatístico bayesiano calcula as respetivas probabilidades *a posteriori*,

$$P(\Theta_0|x) = \int_{\Theta_0} h(\theta|x) d\theta, \qquad P(\Theta_1|x) = \int_{\Theta_1} h(\theta|x) d\theta,$$

e também, usualmente, o rácio das vantagens *a posteriori* de H_0 sobre H_1, seja, $P(\Theta_0|x)/P(\Theta_1|x)$. Pode talvez dizer-se que no quadro bayesiano o resultado da inferência não é tanto a aceitação ou rejeição da hipótese H_0 – como sucede

na doutrina de Neyman-Pearson – mas mais a alteração das credibilidades atribuídas à hipótese e à alternativa. A análise bayesiana passa muitas vezes pela comparação do rácio das vantagens *a posteriori* com o das vantagens *a priori*,

$$B(x) = \frac{P(\Theta_0|x)/P(\Theta_1|x)}{P(\Theta_0)/P(\Theta_1)}, \tag{1.6}$$

que se designa por fator Bayes a favor de H_0 (ou Θ_0) e que traduz o pendor dos dados x para sustentar H_0. Evidentemente, quanto maior for o fator Bayes maior é o aumento das vantagens *a posteriori* em relação às vantagens *a priori* e maior é, portanto, a sustentação que os dados dão à hipótese H_0. O fator Bayes depende, em geral, da distribuição *a priori* e pode expressar-se como um rácio de verosimilhanças ponderadas pelas densidades *a priori* condicionais a cada hipótese sobre Θ_0 e Θ_1 (*vide* Paulino *et al.*, 2003). Neste sentido não pode dizer-se que o fator Bayes seja uma medida de suporte da hipótese H_0 baseada *apenas* nos dados.

Quando a hipótese conjeturada sobre θ é incisiva ao ponto de se definir como $H_0 : \theta = \theta_0$, a via do fator Bayes ou das chances *a posteriori* exige que a distribuição *a priori* seja condizente com tal conjetura em ordem a evitar uma probabilidade nula, revestindo em geral uma natureza mista. Esta implicação é tomada como natural por bayesianos como Jeffreys, com o argumento de que a distribuição *a priori* deve integrar os juízos probabilísticos inerentes à própria formulação das hipóteses em confronto que, no caso vertente, atribuem uma ordem de importância a θ_0 diferenciada da dos outros valores de θ.

Outros bayesianos como Lindley e Zellner advogaram um procedimento distinto, com uma certa analogia formal com os testes de significância clássicos, na situação em que a própria formulação de hipóteses pontuais não interfere com a distribuição *a priori*. Esse procedimento pode ser mais geralmente descrito pela quantificação da plausibilidade relativa *a posteriori* do valor θ_0 através do cálculo da medida $P = P(\theta \notin R_0(x)|x))$, onde $R_0(x) = \{\theta \in \Theta : h(\theta|x) \geq h(\theta_0|x)\}$ é a menor região de credibilidade HPD que contém θ_0. Valores grandes (pequenos) do nível P de plausibilidade relativa *a posteriori* de H_0 apontam para uma evidência a favor de (contra) essa hipótese.

O instrumento fundamental da abordagem bayesiana e a forma de usar o modelo estatístico conjunto $M = \{f(x|\theta)h(\theta), x \in \mathcal{X}, \theta \in \Theta\}$ na realização de inferências deixam antever que a questão de avaliar a razoabilidade do modelo conjeturado em termos absolutos não tem uma resposta popperiana (refuta-

ção/não refutação) do género daquela garantida pelos testes de ajustamento da metodologia clássica.

A via dos fatores Bayes poderá ser utilizada se for possível estender o modelo M (ou parte dele) a uma família mais vasta que se admite integrar o verdadeiro modelo tido como desconhecido, o que permitirá comparar os modelos dentro dela. De outro modo, poderá recorrer-se a várias medidas de adequabilidade para uma análise relativa do desempenho do modelo de partida no quadro de uma classe a definir de modelos competidores (*vide* Paulino *et al.*, 2003). A relativa insatisfação com estas opções tem levado alguns estatísticos a defender a aplicação da abordagem bayesiana apenas quando o modelo de partida é considerado como não questionável, condição esta que Gillies (2001) apelida de *fixity of the theoretical framework*.

1.4.2 Inferências preditivas

Muitos bayesianos consideram que a inferência não tem de se restringir a proposições sobre parâmetros não observáveis. Afirmam, consequentemente, que as inferências paramétricas possuem inconvenientes na medida em que os valores dos parâmetros poucas vezes são conhecidos e portanto as conclusões a que tais inferências conduzem raramente podem ser confrontadas com a realidade. Para bayesianos como Lindley, é mais fundamental o problema que consiste em partir de um conjunto de observações $(x_1, x_2, \ldots, x_n)$ (ontem) e inferir conclusões, em termos de probabilidade (subjetiva, claro), sobre o conjunto de variáveis ainda não observadas, $(x_{n+1}, x_{n+2}, \ldots, x_{n+M})$ (amanhã).

Para facilitar aqui a exposição faz-se $M = 1$ e consideram-se as $n+1$ variáveis aleatórias $X_1, X_2, \ldots, X_n, X_{n+1}$ i.i.d. dado θ com função densidade $f(x|\theta)$ e o problema consiste em predizer o comportamento da variável aleatória X_{n+1} depois de observar $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. Ao tentar predizer a variável X_{n+1} com densidade amostral $f(x|\theta)$ encontram-se dois tipos de aleatoriedade: (a) o que se prende com o facto de a própria variável ser aleatória; (b) o derivado do desconhecimento do valor de θ. Por exemplo, quando se procede à estimação de θ e se obtém $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ pelo método da máxima verosimilhança, e se escreve, $P(a < X_{n+1} < b|x_1, x_2, \ldots, x_n) \cong \int_a^b f(x|\hat{\theta})dx$, para estimar a probabilidade do acontecimento $a < X_{n+1} < b$ está-se a ignorar a aleatoriedade que envolve a substituição do parâmetro pela estimativa, porquanto ambos os tipos de aleatoriedade devem influenciar o processo preditivo. Este

procedimento de enfiar uma estimativa no lugar do parâmetro na distribuição amostral (*plug-in procedure*) merece, pois, grandes reservas.

Embora a solução clássica do problema da predição tenha muito mais que se lhe diga [veja-se Amaral Turkman (1980)], parece poder afirmar-se que os bayesianos têm uma solução bem mais lúcida. Se dispõem apenas da informação inicial, traduzida pela distribuição *a priori* $h(\theta)$, o instrumento a aplicar é naturalmente a já referida distribuição marginal ou distribuição preditiva *a priori* $f(x)$. Caso mais interessante é aquele em que se observa $x = (X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ e se pretende predizer X_{n+1} independente das observações anteriores dado θ [ou predizer $(X_{n+1}, X_{n+2} \ldots, X_{n+M})$ – o problema não é muito diferente]. Utilizando raciocínio completamente probabilístico tem-se $f(x_{n+1}|x) = \int_\Theta f(x_{n+1}|\theta)h(\theta|x)d\theta$, onde a distribuição *a posteriori* ocupa o lugar da distribuição *a priori* por haver a informação dada pela amostra. Podem determinar-se valores típicos dessa distribuição preditiva, probabilidades de qualquer região no espaço de valores de X_{n+1} ou então valores, $a = a(x)$, $b = b(x)$, para um valor prefixado da probabilidade, $P(a < X_{n+1} < b|x) = \int_a^b f(x_{n+1}|x)dx_{n+1}$, que definem um intervalo de predição (com a caraterística HPD ou não).

1.5 Conclusão

Em resumo, segundo a perspetiva bayesiana:

- No esquema clássico a inferência estatística passa por raciocínios do tipo indutivo, como por exemplo os intervalos de confiança, que não têm interpretação direta em termos de probabilidade. A dificuldade ou impossibilidade de fazer inferências com interpretação direta em termos de probabilidade – já que o parâmetro θ não é considerado uma quantidade aleatória – é duramente criticada por Jaynes.

- No esquema bayesiano verifica-se que todas as inferências são realizadas a partir da aplicação lógica do cálculo de probabilidades. A inferência estatística bayesiana, insista-se, não recorre a resultados que não possam deduzir-se a partir das regras de cálculo de probabilidades e em particular do Teorema de Bayes. Como afirma O'Hagan: "*Probability theory is a completely self-consistent system. Any question of probabilities has*

> *one and only one answer, although there may be many ways to derive it*".

Num plano de referência a posições não extremistas convém recordar declarações como a de Dawid (1985) que, embora confessando nítida preferência pela teoria bayesiana, declara *I believe – and believe I have proved – that no approach to statistical inference, Bayesian or not, can ever be entirely satisfactory.* O que alguns estatísticos hoje defendem não é a opção *unicamente Bayes* de autores como Savage mas a posição eclética partilhada por Wasserman (2004) ao afirmar, em síntese, que para combinar credibilidades *a priori* com dados estão indicados os métodos bayesianos, e para construir procedimentos que garantam bons resultados com a longa repetição das observações há que recorrer a métodos frequencistas.

Capítulo 2

Representação da Informação *A Priori*

A colocação em funcionamento da máquina inferencial exige que o insumo do seu instrumento fundamental esteja devidamente preparado. O 1º ingrediente é o modelo amostral que se supõe acomodar (com maior ou menor incerteza) os dados entretanto obtidos de algum processo experimental ou observacional que interessa analisar. Este modelo abrange um conjunto de aspetos desconhecidos sobre os quais pode haver informação apriorística que interessa incluir na análise, por mais vaga ou significativa que ela seja, o que exige saber como se deve representar ou quantificar.

O processo de representação da informação *a priori*, quantas vezes atribulado pelo envolvimento de elementos de natureza subjetiva a eliciar, vai aqui ser abordado por confinamento a duas situações:

- A primeira é aquela em que não existe informação *a priori* palpável, de natureza quer objetiva quer subjetiva (o frequentemente chamado estado de "ignorância *a priori*" ou em que o conhecimento *a priori* é pouco significativo relativamente à informação amostral (o estado de conhecimento "vago" ou "difuso"). Focar-se-ão alguns dos principais métodos advogados para o efeito conduzindo a distribuições *a priori* minimamente informativas em algum sentido e que, em regra, são apelidadas de **distribuições não-informativas**;

- A segunda está muito ligada à adoção prévia de uma forma distribucional conveniente e à escolha de um membro dessa família que se revele consentâneo com medidas-resumo cuidadosamente eliciadas para a distribuição a determinar - como um exemplo interessante deste processo eliciatório veja-se o problema médico tratado em Paulino, Soares e Neuhaus (2003). É neste âmbito que se inserem as denominadas **distribuições conjugadas naturais**, as quais podem igualmente servir como geradoras de distribuições não-informativas impróprias.

Para informação adicional sobre o problema de escolha da distribuição *a priori* (outros métodos geradores de distribuições vagas ou de eliciação de distribuições subjetivas, *vide* O'Hagan (2010), Kass e Wasserman (1996) e Paulino, Amaral Turkman e Murteira (2003).

2.1 Distribuições não-informativas

Estas distribuições começaram por ser dominantemente interpretadas como representações formais de ignorância, mas há hoje uma tendência (motivada pela não aceitação de representações objetivas únicas da ignorância) para encará-las como opções convencionais de defeito a que se recorre em caso de informação *a priori* insuficiente que torne difícil eliciar uma distribuição subjetiva considerada adequada. Independentemente da interpretação, este tipo de distribuições pode desempenhar ainda um papel de referência, mesmo que se disponha de fortes crenças *a priori*, como forma de:

- deduzir as crenças *a posteriori* para quem parte de um conhecimento escasso (*i.e.*, quando a amostra fornece o grosso da informação sobre o parâmetro) e, nessa medida, se acha incapaz de determinar subjetivamente uma distribuição razoável – assim se reconhecendo a sua própria ignorância;

- permitir a comparação com os resultados da inferência clássica que "só" usa a informação amostral (no todo ou em parte);

- averiguar a influência nas inferências da distribuição *a priori* subjetiva que descreve a informação realmente existente, quando confrontada com as que resultam do uso da distribuição *a priori* de referência.

Descrevem-se sucintamente em seguida alguns dos argumentos mais usados que conduzem a este tipo de distribuições.

Método de Bayes-Laplace

Este método baseia-se na invocação do Princípio da Razão Insuficiente decorrente da escassez informativa *a priori* para adotar a ideia de equiprobabilidade. Dependendo da cardinalidade de Θ este argumento conduz às distribuições Uniforme discreta ou Uniforme contínua.

Se no caso de o número de valores de θ ser finito, e.g. $\Theta = \{\theta_1, \ldots, \theta_k\}$, este argumento pode ser considerado pacífico ao levar à distribuição $h(\theta) = 1/k,\ \theta \in \Theta$, o mesmo não sucede nas outras situações. Com efeito, se Θ for infinito numerável, a distribuição decorrente é imprópria, o mesmo sucedendo no caso de Θ ser um conjunto infinito não numerável e não limitado, o que causa algum desconforto a estatísticos que não simpatizam com medidas não normalizadas (ainda que tal não ponha necessariamente em causa o seu uso no teorema de Bayes, pois a distribuição *a posteriori* – o fulcro das inferências – é muitas vezes própria).

Uma outra crítica, talvez mais séria, ao uso do argumento de que a ausência de informação, que alguns rotulam de ignorância, deve ser representada por uma distribuição uniforme, resulta do facto de esta não ser invariante relativamente a transformações não lineares, conduzindo assim a contrassensos probabilísticos. Como ilustração, tome-se o modelo $\{Ber(\theta),\ \theta \in (0,1)\}$ que faz parte da família exponencial com parâmetro natural $\psi = \ln[\theta/(1-\theta)] \in \mathbb{R}$. O uso simultâneo de distribuições uniformes para θ (própria) e ψ (imprópria) é probabilisticamente inconsistente já que $\theta \sim U(]0,1[) \equiv Be(1,1)$ equivale para ψ a uma distribuição logística reduzida, de função densidade $h(\psi) = \frac{e^{\psi}}{(1+e^{\psi})^2},\ \psi \in \mathbb{R}$.

No caso geral, sendo $\psi = \psi(\theta)$ uma tranformação injetiva de um parâmetro θ, que assume uma gama contínua de valores possíveis, e $h(\theta)$ uma densidade *a priori* para θ, então

$$h(\psi) = h\left[\theta(\psi)\right] \left| \frac{d\theta}{d\psi} \right| \tag{2.1}$$

deve ser a correspondente densidade para a reparametrização ψ, que não é uniforme quando $h(\theta)$ o é se o jacobiano depende de ψ, como acontece com transformações não lineares do género da referida no exemplo acima.

Método de Jeffreys

Entre os procedimentos que asseguram invariância sob transformações injetivas está aquele advogado por Jeffreys e que se baseia no uso da medida de informação de Fisher sobre $\theta \in \mathbb{R}$, definida por

$$I(\theta) = E\left[\left(\frac{\partial \ln f(X \mid \theta)}{\partial \theta}\right)^2 \middle| \theta\right].$$

Com efeito, o facto de para qualquer transformação real injetiva de $\theta \in \mathbb{R}$ se ter

$$I(\psi) = I(\theta(\psi))\left(\frac{d\theta}{d\psi}\right)^2$$

mostra que a distribuição proposta por Jeffreys para o caso uniparamétrico, $h(\theta) \propto [I(\theta)]^{\frac{1}{2}}$, goza da referida propriedade de invariância e, deste modo, assegura a identidade das inferências qualquer que seja a transformação biunívoca no espaço paramétrico (ou seja, reparametrização) que se use.

Além disso, deve notar-se que $I(\theta)$ é tanto maior quanto maior for a taxa quadrática de variação (em média ao longo do espaço amostral) com θ de $\ln f(X \mid \theta)$, i.e., quanto mais diferenciado pelo modelo estiver θ de $\theta + d\theta$. Considerar então mais (menos) plausíveis *a priori* os valores de θ com maior (menor) $I(\theta)$, *i.e.*, sobre os quais há maior (menor) informação amostral, corresponde a considerar reduzido tanto quanto possível o efeito da informação *a priori*. Daí o carácter não informativo das distribuições *a priori* geradas pela regra de Jeffreys. O seu carácter alegadamente objetivo advém de serem automaticamente obtidas do próprio modelo supostamente gerador dos dados.

Exemplo 2.1 Considere-se um modelo amostral cuja função de verosimilhança se pode escrever como

$$L(\theta|x,n) = k\,\theta^x(1-\theta)^{n-x}, \quad \theta \in (0,1),$$

onde k não depende de θ. Se n é fixo e $k = \binom{n}{x}$ o modelo corresponde a $X|n,\theta \sim Bi(n,\theta)$, cujo valor médio é $n\theta$ donde $I(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, pelo que a distribuição de Jeffreys é própria e definida por

$$h(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2},\ \theta \in (0,1) \Leftrightarrow \theta \sim Be(1/2,1/2),$$

correspondendo por aplicação do argumento de transformação de variáveis à distribuição $U(0,2\pi)$ para $\psi = \text{arc sen}\sqrt{\theta}$.

Se alternativamente x é fixo e $k = \binom{n-1}{x-1}$ está-se na presença do modelo $N - x|x, \theta \sim BiN(x, \theta)$, de valor médio $x(1-\theta)/\theta$ donde $I(\theta) \propto \theta^{-2}(1-\theta)^{-1}$, implicando que a distribuição de Jeffreys é tal que

$$h(\theta) \propto \theta^{-1}(1-\theta)^{-1/2},\ \theta \in (0,1),$$

correspondendo a uma distribuição imprópria que se simboliza por "$Be(0, 1/2)$" e que é consistente com uma distribuição "Uniforme" para a função

$$\psi = \ln \frac{1-\sqrt{1-\theta}}{1+\sqrt{1-\theta}}.$$

■

Este exemplo evidencia o que era expectável, que a distribuição de Jeffreys pela sua definição depende plenamente da distribuição amostral e não apenas do seu núcleo. Esta dependência para com o espaço amostral é para uns motivo de veemente crítica, na base nomeadamente de conduzir a inferências *a posteriori* sobre o mesmo parâmetro distintas consoante a natureza da experiência efetivamente realizada (no exemplo, amostragem binomial direta ou binomial inversa), ainda que a diferença possa ser bem ligeira em amostras moderadas, como no exemplo ilustrativo em questão. Para outros, tal dependência é considerada legítima pelo argumento de que a vaguidade da informação *a priori*, que a distribuição de Jeffreys visa representar, deve ser vista em função da informação associada ao tipo de amostragem planeada e não em termos absolutos.

A aplicação da regra de Jeffreys a modelos uniparamétricos de localização, $\{f(x|\theta) = g(x-\theta),\ \theta \in \Theta \subseteq I\!R\}$ (e.g., modelo Normal com variância conhecida), conduz à distribuição Uniforme contínua que é invariante sob transformações lineares (i.e., face a translações) e imprópria se Θ for um conjunto não limitado. Se for aplicada a modelos uniparamétricos de escala, $\{f(x|\theta) = \frac{1}{\theta}g(x/\theta),\ \theta \in \Theta \subseteq I\!R_+\}$ (e.g., modelo Normal com média conhecida), conduz à distribuição imprópria $h(\theta) \propto \theta^{-1} I_{(0,+\infty)}(\theta)$, que é invariante sob potências (i.e., face a transformações de escala).

Em modelos multiparamétricos a regra de Jeffreys é baseada na raiz quadrada do determinante da matriz de informação de Fisher. Contudo, devido a implicações *a posteriori* indesejáveis, esta regra costuma ser preterida, até por sugestão do próprio Jeffreys, pela imposição prévia de independência *a*

priori entre parâmetros (particularmente quando são de natureza diferente) e pelo uso das regras de Jeffreys uniparamétricas para a especificação das distribuições marginais. Por exemplo, no modelo $\{N(\mu, \sigma^2) : \mu \in I\!R, \sigma^2 \in I\!R_+\}$ a distribuição de Jeffreys é dada por

$$h(\mu, \sigma^2) \propto \sigma^{-m},\ \mu \in I\!R,\ \sigma^2 > 0,$$

com $m = 3$ segundo a regra biparamétrica e $m = 2$ pela regra assente na referida fatorização da distribuição conjunta dos parâmetros de localização, μ, e de escala, σ^2.

Método da entropia máxima

A noção de entropia emprestada da Física, ao estar associada a uma medida de incerteza, foi sugerida por Jaynes como um meio de chegar a uma distribuição *a priori* que possa representar um estado de ignorância pelo menos relativa. Tal distribuição deveria então corresponder ao valor máximo possível da entropia.

Tomando entropia de uma distribuição $h(\theta),\ \theta \in \Theta$ como o valor esperado $\mathcal{E}(h(\theta)) = E_h[-\ln h(\theta)]$, facilmente se mostra no caso finito quando $\Theta = \{\theta_1, \ldots, \theta_k\}$ que a distribuição de entropia máxima (i.e., com a maior incerteza) é a Uniforme discreta $h(\theta_i) = 1/k,\ i = 1, \ldots, k$, à qual corresponde a entropia de $\ln k$. Basta maximizar a função lagrangiana definida por $\mathcal{E}(h(\theta))$ acrescida do termo $\lambda\left(\sum_{i=1}^{k} h(\theta_i) - 1\right)$, onde λ representa o multiplicador de Lagrange associado à restrição natural de uma função de probabilidade.

Se se pretender maximizar a entropia restringida à informação representada por valores preespecificados de momentos ou quantis, e.g., na forma $E(g_j(\theta)) = u_j,\ j = 1, \ldots, m$, basta usar a mesma via (método dos multiplicadores de Lagrange) com introdução de tais restrições adicionais para se chegar à expressão

$$h(\theta_i) = \frac{\exp\{\sum_{j=1}^{m} \lambda_j g_j(\theta_i)\}}{\sum_{l=1}^{k} \exp\{\sum_{j=1}^{m} \lambda_j g_j(\theta_l)\}},$$

onde os valores dos m multiplicadores λ_j se obtêm usando as correspondentes restrições introduzidas.

Exemplo 2.2 No contexto discreto que vem sendo considerado, admita-se $\theta_i = i$ pelo que $\Theta = \{1, \ldots, k\}$, e que a mediana é um dos valores possíveis

identificado por q. Assim, tem-se uma restrição imposta com $u_1 = q$ e $g_1(\theta)$ a função indicadora do acontecimento $\theta \leq q$, ou seja, dada por $\sum_{i=1}^{q} h(i) = 1/2$. Pela expressão acima

$$h(i) = \begin{cases} \frac{e^{\lambda_1}}{e^{\lambda_1} q + (k-q)}, & \text{se } i \leq q \\ \frac{1}{e^{\lambda_1} q + (k-q)}, & \text{se } q < i \leq k \end{cases},$$

onde $e^{\lambda_1} = (k-q)/q$ pela restrição identificadora do valor da mediana. Obtém-se então

$$h(i) = \begin{cases} \frac{1}{2q}, & \text{se } i \leq q \\ \frac{1}{2(k-q)}, & \text{se } q < i \leq k \end{cases},$$

ou seja, uma distribuição uniforme por troços. ■

No caso contínuo em que Θ é um intervalo limitado da reta real, o recurso ao Cálculo de Variações permite mostrar que a distribuição de entropia máxima é Uniforme contínua que, como se sabe já, não é invariante sob toda a transformação injetiva, o que cria problemas à visão da entropia $\mathcal{E}(h(\theta))$ como uma medida absoluta de incerteza.

Baseando-se na relação entre entropia e e medida de informação de Kullback-Leibler no caso discreto, Jaynes (1968) redefine entropia no caso contínuo em relação a uma distribuição de referência não informativa $h_0(\theta)$ como $\mathcal{E}(h(\theta)) = E_h\left[-\ln \frac{h(\theta)}{h_0(\theta)}\right]$.

Se se supuser a existência de informação inicial representada por restrições como se definiram acima, o uso de novo do cálculo variacional leva a que a solução do problema de maximização seja expressável por

$$h(\theta) \propto h_0(\theta) \exp\left\{\sum_{j=1}^{m} \lambda_j \; g_j(\theta)\right\},$$

onde os multiplicadores λ_j se obtêm a partir das restrições consideradas.

Exemplo 2.3 Considere-se que θ é um parâmetro de localização mas que se sabe ser positivo pelo que $\Theta = (0, +\infty)$ e que o seu valor médio é u. Adotando-se como distribuição *a priori* não informativa invariante sob translações a distribuição "Uniforme" em Θ, tem-se $h(\theta) \propto \exp(\lambda_1 \theta), \; \theta > 0$, o que implica pela restrição natural que $h(\theta) = -\lambda_1 \exp(\lambda_1 \theta) I_{(0,+\infty)}(\theta)$, com $\lambda_1 < 0$, ou seja, uma distribuição Exponencial. Tendo em conta que o seu valor médio

prefixado é $-1/\lambda_1 = u$, segue-se que a distribuição de entropia máxima é $\theta \sim Exp(1/u)$. ■

Exemplo 2.4 Seja θ novamente um parâmetro de localização tal que $\Theta = I\!R$ e suponha-se que $E(\theta) = u_1$ e $Var(\theta) = u_2$. Deste modo, usando a mesma distribuição de referência (imprópria) referida no exemplo anterior, tem-se que $h(\theta) \propto \exp\{\lambda_1\theta + \lambda_2(\theta - u_1)^2\}$, $\theta \in I\!R$, onde por simples manipulação algébrica se pode reescrever

$$\lambda_1\theta + \lambda_2(\theta - u_1)^2 = \lambda_2\left[\theta - \left(u_1 - \frac{\lambda_1}{2\lambda_2}\right)\right]^2 + \left[\lambda_1 u_1 - \frac{\lambda_1^2}{4\lambda_2}\right].$$

Consequentemente,

$$h(\theta) \propto \exp\left\{\lambda_2\left[\theta - \left(u_1 - \frac{\lambda_1}{2\lambda_2}\right)\right]^2\right\},$$

que corresponde ao núcleo de uma distribuição gaussiana de valor médio $u_1 - \lambda_1/(2\lambda_2)$ e variância $-1/(2\lambda_2)$ com $\lambda_2 < 0$. Atendendo aos dois momentos preespecificados conclui-se que $\lambda_1 = 0$ e $\lambda_2 = -1/(2u_2)$, conduzindo à distribuição *a priori* de entropia máxima $\theta \sim N(u_1, u_2)$. ■

2.2 Distribuições conjugadas naturais

A família distribucional com uma dada estrutura funcional a selecionar, onde se vai procurar um membro condizente com os resumos eliciados, deve idealmente satisfazer os seguintes requisitos:

- Versatilidade para acomodar o maior número possível de crenças *a priori*;
- Acessibilidade interpretativa para facilitar o processo de sumariação dos seus membros;
- Simplicidade da derivação analítica das distribuições *a posteriori* e preditivas.

A simplicidade da operação bayesiana poderá ficar garantida se se impuser que a família de distribuições *a priori* $\mathcal{H} = \{h(\theta \mid a) : a \in \mathcal{A}\}$, onde $\mathcal{A}$ denota o conjunto de valores para os índices rotuladores das várias distribuições – os

denominados hiperparâmetros – seja **fechada sob amostragem** de (qualquer elemento de) $\mathcal{F} = \{f(x \mid \theta) : \theta \in \Theta\}$, i.e., que

$$h(\theta) \in \mathcal{H} \Rightarrow h(\theta \mid x) \propto h(\theta) f(x \mid \theta) \in \mathcal{H}.$$

Nestas condições, diz-se também que $\mathcal{H}$ é uma **família conjugada natural** de $\mathcal{F}$. De outra forma, a família $\mathcal{H}$ diz-se conjugada natural de $\mathcal{F}$ se $L(\theta \mid x) \equiv f(x \mid \theta)$, para cada x, é proporcional a um membro de $\mathcal{H}$ e $\mathcal{H}$ é fechada em relação a produtos, i.e., para todo o a_0, $a_1 \in \mathcal{A}$, existe $a_2 \in \mathcal{A}$ tal que

$$h(\theta \mid a_0) h(\theta \mid a_1) \propto h(\theta \mid a_2).$$

Exemplo 2.5 Sendo $x = (x_i,\ i = 1, \ldots, n)$ uma concretização de uma amostra aleatória do modelo bernoulliano $Ber(\theta)$, tem-se

$$f(x_1, \ldots, x_n | \theta) = \theta^{\sum_i x_i} (1-\theta)^{n - \sum_i x_i},$$

que é proporcional ao núcleo de uma distribuição $Be(\sum_i x_i + 1, n - \sum_i x_i + 1)$ para θ que é fechado relativamente a produtos. A família conjugada natural do modelo amostral de Bernoulli é então a família Beta, cuja versatilidade é bem conhecida. Por conseguinte,

$$\theta \sim Be(a, b) \Rightarrow \theta | x \sim Be(A, B), \quad A = a + \sum_i x_i, B = b + n - \sum_i x_i,$$

revelando como a informação amostral se materializa na facilmente derivável distribuição *a posteriori* através dos números de sucessos e de insucessos (ou seja, da estatística suficiente mínima), conjugados aditivamente com os hiperparâmetros representantes da informação *a priori*.

Como $a, b > 0$, a informação *a priori* anula-se em termos relativos fazendo $a, b \to 0$, pelo que a distribuição não-informativa (ou vaga) obtida da família conjugada natural é a distribuição imprópria de Haldane, "$Be(0,0)$", definida por $h(\theta) \propto \theta^{-1}(1-\theta)^{-1},\ \theta \in (0,1)$, que corresponde à distribuição "Uniforme" para $\psi = \ln[\theta/(1-\theta)] \in I\!R$. Na sequência, a distribuição *a priori* $Be(a, b)$ é interpretável como a distribuição *a posteriori* resultante da atualização dessa distribuição não-informativa por uma pseudoamostra de tamanho $a + b$ com a sucessos. ■

O facto de a operação bayesiana de conjugação das informações apriorística e amostral na família conjugada natural se processar dentro dela conduz a que

possa ser simbolicamente representada por uma transformação no espaço $\mathcal{A}$ dos hiperparâmetros

$$a \in \mathcal{A} \stackrel{\mathcal{F}}{\to} A = a + (A - a) \in \mathcal{A}.$$

Esta transformação, formatável de modo a evidenciar os pesos relativos dos dois tipos de informação, acaba por consubstanciar em si a simplicidade interpretativa e analítica do mecanismo bayesiano no quadro da família conjugada natural. Na forma acima, $A - a$ exprime a influência da informação amostral na alteração da informação *a priori* materializada em a e a ilustração disso conseguiu-se claramente com o exemplo anterior.

Exemplo 2.6 Se no exemplo anterior a amostra aleatória fosse respeitante ao modelo $Geo(\theta)$ de função de probabilidade $f(x_i|\theta) = \theta(1-\theta)^{x_i},\ x_i \in I\!N_0$, a aplicação do mesmo raciocínio conduziria à mesma família conjugada natural e à mesma distribuição não-informativa. Todavia, ter-se-ia $\theta|x \sim Be(A, B), \quad A = a + n, B = b + \sum_i x_i$ para uma distribuição *a priori* $Be(a, b)$ que seria então visualizável como uma distribuição *a posteriori* resultante da atualização de "$Be(0, 0)$" por uma amostra fictícia do modelo Geométrico de tamanho a e número total de insucessos b. ■

Exemplo 2.7 Sendo $x = (x_i,\ i = 1, \ldots, n)$ uma concretização de uma amostra aleatória do modelo Erlang $Ga(m, \lambda)$, onde $m \in I\!N$ é suposto conhecido, a função densidade amostral é tal que $f(x_1, \ldots, x_n|\lambda) \propto \lambda^{mn} e^{-\lambda \sum_i x_i}$, apresentando um núcleo $Ga(mn + 1, \sum_i x_i)$ para λ que é fechado relativamente a produtos. Consequentemente, a família Gama é a conjugada natural sob amostragem do modelo Erlang, sendo então $\lambda|x \sim Ga(A, B)$, com $A = a+mn$, $B = b+\sum_i x_i$, a distribuição *a posteriori* correspondente à distribuição *a priori* $Ga(a, b),\ a, b > 0$. Esta distribuição *a priori* é então interpretável como resultante da atualização da distribuição vaga "$Ga(0, 0)$", definida por $h(\lambda) \propto \lambda^{-1},\ \lambda > 0$, por uma pseudoamostra erlangiana de tamanho a/m e média empírica das respetivas observações mb/a. ■

Exemplo 2.8 Considerando-se agora uma amostra aleatória do modelo Normal de valor médio μ e precisão $1/\sigma^2$ conhecida, o núcleo da correspondente função densidade numa concretização daquela amostra $x = (x_i,\ i = 1, \ldots, n)$ pode escrever-se como

$$f(x_1, \ldots, x_n|\mu) \propto e^{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2},$$

mostrando que é proporcional ao núcleo de uma distribuição gaussiana para μ de valor médio $\bar{x}$ e variância σ^2/n conhecida. O produto de dois núcleos deste tipo ainda é um núcleo do mesmo tipo[6].

Deste modo, a família conjugada natural é gaussiana, verificando-se $\mu \sim N(a, b^2) \Rightarrow \mu|x \sim N(A, B^2)$, onde pela identidade mencionada

$$A = \frac{\frac{1}{b^2}a + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{b^2} + \frac{n}{\sigma^2}}, \quad B^2 = \left(\frac{1}{b^2} + \frac{n}{\sigma^2}\right)^{-1}.$$

Fazendo $b \to +\infty$ obtém-se a distribuição "Uniforme" em $\mathbb{R}$ como distribuição vaga, a qual por sua vez implica a distribuição *a posteriori* $\mu|x \sim N(\bar{x}, \sigma^2/n)$. Em consonância, a distribuição *a priori* $\mu \sim N(a, b^2)$ pode ser encarada como fruto da atualização da referida distribuição vaga por uma amostra fictícia de tamanho m do modelo Normal, com média empírica a e variância conhecida mb^2. ■

Apesar de os exemplos ilustrativos neste capítulo se reportarem a modelos amostrais uniparamétricos, o raciocínio para identificar a família conjugada natural de modelos multiparamétricos, se existente, mantém-se integralmente como aliás se patenteia em exemplos descritos no capítulo que se segue a este.

A grande diferença reside na incomparavelmente maior dificuldade em eliciar o maior número de resumos necessários para identificar a distribuição conjugada natural num contexto multivariado que, infelizmente, não possui a mesma diversidade de formas distribucionais do caso univariado. Estratégias que rodeiam ou ultrapassam estes reveses na escolha de uma distribuição *a priori* fiável incluem a especificação de independência entre subconjuntos de parâmetros e a adoção de misturas contínuas ou finitas de distribuições conjugadas naturais.

Também outros métodos particularmente delineados para modelos multiparamétricos poderão ser usados para representar a distribuição *a priori*, em especial, o método das distribuições objetivas de referência devido a Berger e Bernardo e que se encontra detalhadamente descrito em Bernardo e Smith (1994). Este método nem sempre é bem-sucedido pela impossibilidade de obter analiticamente os resultados pretendidos, mormente em modelos parametricamente complexos.

[6] Atenda-se à identidade algébrica $d_1(z-c_1)^2 + d_2(z-c_2)^2 = (d_1+d_2)(z-c)^2 + \frac{d_1 d_2}{d_1+d_2}(c_1 - c_2)^2$, onde $c = \frac{d_1 c_1 + d_2 c_2}{d_1+d_2}$.

Capítulo 3

Metodologia Bayesiana em Aplicações Básicas

Tendo conhecimento das ideias fundamentais da abordagem bayesiana à Inferência Estatística e os procedimentos mais relevantes para a representação da informação *a priori* importa passar-se para uma fase ilustrativa. Julga-se ser incontroverso que a ilustração básica do paradigma bayesiano deva reportar-se à sua aplicação a situações onde os resultados inferencialmente relevantes possam ser deduzidos em termos tanto quanto possível exatos, de preferência analiticamente ou quando muito através de simulação operada através de distribuições *a posteriori* perfeitamente conhecidas.

É isso que vai ser feito neste capítulo ao proceder-se à análise de uma dezena de modelos bayesianos capazes de conduzir a uma resolução potencialmente apropriada de correspondentes exercícios maioritariamente do âmbito de cursos básicos de Estatística. Sendo os modelos bayesianos definidos como a distribuição conjunta do vetor (X) de observações e dos parâmetros (θ), a sua descrição de partida vai basear-se no modelo amostral $\{f(x|\theta)\}$ (o modelo da denominada Estatística Clássica) e na distribuição *a priori* $h(\theta)$, que se indicará notacionalmente por algo como $f(x|\theta) \wedge h(\theta)$ em forma predominantemente textual.

3.1 Modelo Binomial ∧ Beta

Seja $x = (x_i,\ i = 1, \ldots, n)$ uma concretização das variáveis aleatórias condicionalmente independentes $X_i|m_i, \theta \sim Bi(m_i, \theta),\ i = 1, \ldots, n$, com m_i conhecidos e o parâmetro desconhecido θ munido de uma distribuição *a priori* $Be(a, b)$ com os respetivos hiperparâmetros fixados. Então a função densidade *a posteriori* de θ apresenta o núcleo

$$\begin{aligned} h(\theta|x) &\propto \prod_{i=1}^{n} \left\{ \binom{m_i}{x_i} \theta^{x_i} (1-\theta)^{m_i - x_i} \right\} h(\theta|a, b) \\ &\propto \theta^{a + \sum_i x_i - 1} (1-\theta)^{b + \sum_i (m_i - x_i) - 1}, \end{aligned}$$

evidenciando que $\theta|x \sim Be(A, B)$, em que $A = a + \sum_i x_i$ e $B = b + \sum_i (m_i - x_i)$.

Esta distribuição *a posteriori* equivale a usar para transformações de θ de interesse em certas aplicações outros tipos distribucionais como as distribuições F e Z de Fisher,

$$(B/A)\frac{\theta}{1-\theta}\bigg| x \sim F_{(2A,2B)};\ \left[(1/2)\ln(B/A) + (1/2)\ln\frac{\theta}{1-\theta} \right] \bigg| x \sim Z_{(2A,2B)}.$$

Momentos mistos da distribuição *a posteriori* de θ são facilmente derivados como

$$E\left[\theta^{r_1}(1-\theta)^{r_2}|x\right] = \frac{B(A + r_1, B + r_2)}{B(A, B)},$$

o que permite calcular vários resumos dessa distribuição como a média, $E(\theta|x) = A/(A + B)$ e a variância *a posteriori*. Outra estimativa pontual relevante, a moda *a posteriori*, é bem definida quando $A, B > 1$ por $m_0 = \frac{A-1}{A+B-2}$. Quantis e probabilidades *a posteriori* de θ podem ser calculados de funções beta incompletas (que não têm uma clara expressão explícita), ou no caso de a, b inteiros, de funções de distribuição binomial atendendo a que

$$F_{Be(A,B)}(\theta_0) = 1 - F_{Bi(A+B-1,\theta_0)}(A - 1).$$

Tendo em vista a realização de predições, considere-se novas v. a. independentes das já observadas, $Y_j \equiv X_{n+j}, j = 1, \ldots, k \underset{iid|\theta}{\sim} Bi(m_{n+j}, \theta)$. Como por definição a distribuição preditiva *a posteriori* de $Y = (Y_1, \ldots, Y_k)$ é a mistura da sua distribuição amostral (produto de Binomiais) pela distribuição *a*

posteriori (Beta) de θ, a sua função de probabilidade preditiva é expressável por

$$p(y_1,\ldots,y_k|x) = \left[\prod_{j=1}^{k}\binom{m_{n+j}}{y_j}\right]\frac{B(A+y_{\cdot},B+m_{n+\cdot}-y_{\cdot})}{B(A,B)},$$

em que $y_{\cdot} = \sum_{j=1}^{k} y_j$, e $m_{n+\cdot} = \sum_j m_{n+j}$, podendo ser vantajosamente definida como o produto de duas funções de probabilidade, uma condicional relativa a uma distribuição Hipergeométrica multivariada e a outra marginal *a posteriori* de tipo Binomial-Beta,

$$p(y_1,\ldots,y_k|x) = p_{Hpg(\{m_{n+j}\},y_{\cdot})}(y_1,\ldots,y_k|y_{\cdot}) \times p_{BiBe(m_{n+\cdot},A,B)}(y_{\cdot}|x).$$

Com efeito, esta forma elucida o tipo de dependência existente na distribuição preditiva multivariada, bem como a sua natureza univariada de tipo $BiBe(m_{n+1},A,B)$ quando k=1, de valor médio $E(Y_1|x) = m_{n+1}\frac{A}{A+B}$ e variância $Var(Y_1|x) = m_{n+1}\frac{AB}{(A+B)(A+B+1)}(1+\frac{m_{n+1}}{A+B})$.

3.2 Modelo Poisson ∧ Gama

Considere-se agora que $x = (x_i,\ i = 1,\ldots,n)$ é uma concretização de uma amostra aleatória do modelo $Poi(\theta)$ de modo que a correspondente função de probabilidade amostral $f(x_1,\ldots,x_n|\theta) \propto e^{-n\theta}\theta^{x_{\cdot}}$. Como esta é proporcional ao núcleo de uma $Ga(x_{\cdot}+1,n)$ para θ, que é fechado sob produtos, fica claro que a família conjugada natural é a família Gama pelo que

$$\theta \sim Ga(a,b) \ \Rightarrow\ \theta|x \sim Ga(A,B),\ A = a+x_{\cdot}, B = b+n \ \Leftrightarrow\ \alpha = 2B\theta \sim \chi^2_{2A},$$

sendo $h(\theta) \propto \theta^{-1}I_{(0,+\infty)}(\theta)$ a decorrente distribuição difusa imprópria.

Probabilidades *a posteriori* de acontecimentos prefixados em $\Theta = I\!R_+$ ou níveis de plausibilidade relativa *a posteriori* de hipóteses pontuais sobre θ são calculáveis de funções gama incompletas ou de funções de distribuição Qui-quadrado. Momentos *a posteriori* são dados por

$$E(\theta^r|x) = \frac{\Gamma(A+r)}{\Gamma(A)}\frac{B^A}{B^{A+r}}.$$

Para efeitos de predição, sejam $Y_j \equiv X_{n+j}, j = 1,\ldots,k \overset{iid|\theta}{\sim} Poi(\theta)$, independentes da amostra aleatória observada. A distribuição preditiva *a posteriori*

de $Y = (Y_1, \ldots, Y_k)$ é a mistura da sua distribuição amostral (produto de distribuições Poisson) pela distribuição *a posteriori* $Ga(A, B)$, com a respetiva função de probabilidade preditiva definida após integração em θ por

$$p(y_1, \ldots, y_k|x) = \frac{\Gamma(A + y_\bullet)}{\Gamma(A) \prod_j y_j!} \left(\frac{B}{B+k}\right)^A \left(\frac{1}{B+k}\right)^{y_\bullet}.$$

Como a distribuição amostral de $Y_\bullet = \sum_j Y_j$ é $Poi(k\theta)$, um cálculo análogo ao anterior mostra que a distribuição preditiva *a posteriori* de $Y_\bullet$ é a mistura Poi-Ga, de função de probabilidade expressa por $p(y_1, \ldots, y_k|x) \prod_j y_j!(k^{y_\bullet})/y_\bullet!$, mais conhecida por distribuição Binomial Negativa generalizada com parâmetros $(A, B/(B+k))$ traduzindo o número fixado de "sucessos" e a probabilidade de ocorrência de cada um deles. Deste modo, a função de probabilidade preditiva *a posteriori* de $(Y_1, \ldots, Y_k)$ pode visualizar-se esclarecedoramente por

$$p(y_1, \ldots, y_k|x) = p_{M_{k-1}(y_\bullet, \frac{1}{k}\mathbf{1_k})}(y_1, \ldots, y_k|y_\bullet) \times p_{BiN(A,B/(B+k))}(y_\bullet|x),$$

como o produto de uma função de probabilidade condicional Multinomial homogénea por uma função de probabilidade marginal *a posteriori* Poi-Ga. Esta representação ilustra o tipo de dependência no seio da distribuição preditiva *a posteriori* que é do tipo Binomial Negativa quando k=1.

3.3 Modelo Rayleigh ∧ Gama

A distribuição Rayleigh com função densidade $f(x|\delta) = \delta\, x\, e^{-\delta x^2/2}\, I_{(0,+\infty)}(x)$ é um modelo probabilístico relevante em alguns problemas de Engenharia. Se $x = (x_i,\ i = 1, \ldots, n)$ for uma concretização de uma amostra aleatória desse modelo amostral e se se usar uma distribuição *a priori* $\delta \sim Ga(a, b)$, facilmente se conclui que $\delta|x \sim Ga(A, B),\ A = a + n, B = b + \sum_i x_i^2/2$. Seria previsível que a família Gama fosse a conjugada natural do modelo Rayleigh já que o núcleo da densidade deste é o de uma distribuição Gama que é fechado perante produtos.

Resumos da distribuição *a posteriori* Gama de δ podem ser obtidos da expressão dos seus momentos simples referida na secção anterior. Em especial, $E(\delta|x) = A/B$ e $Var(\delta|x) = A/B^2$. Uma outra estimativa bayesiana de δ é a moda *a posteriori* dada por (A-1)/B quando $A > 1$.

Testes de hipóteses $H_0 : \delta = \delta_0$ formuladas sem impacte na distribuição *a priori* podem ser efetuados recorrendo a funções gama incompletas. Por

exemplo, usando a distribuição *a priori* de Jeffreys $h(\delta) \propto \delta^{-1} I_{(0,+\infty)}(\delta)$, a distribuição *a posteriori* associada a uma observação x do modelo Rayleigh é $\delta|x \sim Exp(x^2/2)$ pelo que o nível de plausibilidade relativa *a posteriori* de H_0 é $P = P(\delta \geq \delta_0|x) = e^{-\delta_0 x^2/2}$.

Suponha-se agora que se pretendia testar H_0 com $\delta_0 = 2$ na base da informação *a priori* de quem conjeturou tal hipótese e que crê na veracidade dela com uma probabilidade de 50%, adotando para δ sob a hipótese alternativa uma distribuição $Ga(0,02; 0,01)$, e numa amostra de tamanho 5 de medições do modelo Rayleigh tal que $\sum_i x_i^2 = 7,54$. Note-se que a distribuição *a priori* para $\delta|H_1$ é uma distribuição própria de valor médio 2 e variância 200, razoavelmente próxima da distribuição imprópria de Jeffreys ao ter uma densidade $h_1(\delta)$ bastante achatada (ou plana) na maior parte do seu suporte, à exceção da gama de menores valores positivos. De acordo com a informação referida, o fator Bayes a favor de H_0 é bem favorável a esta já que

$$B(x) = \frac{P(H_0|x)}{P(H_1|x)} = \frac{f(x|\delta = 2)}{\int_{\delta \neq 2} f(x|\delta) h_1(\delta) d\delta} = 2^5 e^{-7,54} \frac{\Gamma(0,02) 3,78^{5,02}}{\Gamma(5,02) 0,01^{0,02}} \approx 29,53.$$

Pretendendo-se predizer Y tal que $Y|\delta \sim Ray(\delta)$ independentemente de $(X_1, \ldots, X_n)$, facilmente se conclui que a sua função densidade preditiva *a posteriori* é definida por

$$p(y|x) = AB^A y (B + y^2/2)^{-(A+1)} I_{(0,+\infty)}(y).$$

Resumos desta distribuição preditiva *a posteriori* obtêm-se sem dificuldade. Por exemplo, a predição pontual média é $E(Y|x) = \sqrt{\frac{B}{2}}\, B(1/2, A - 1/2)$ (aplique-se a propriedade sequencial da esperança condicional e prove-se via integração por partes que $E(Y|\delta) = \sqrt{\pi/2}\ \delta^{-1/2}$). Probabilidades preditivas *a posteriori* são de cálculo direto, e.g. $P(Y > 1|x) = (\frac{B}{B+1/2})^A$.

3.4 Modelo Uniforme ∧ Pareto

Sendo $x = (x_i,\ i = 1, \ldots, n)$ uma concretização de uma amostra aleatória do modelo $U([0, \theta])$, a respetiva função densidade amostral

$$f(x_1, \ldots, x_n|\theta) = \theta^{-n} I_{[t,+\infty)}(\theta),\ t = x_{(n)} \equiv \max_{1 \leq i \leq n} x_i$$

corresponde ao núcleo de uma distribuição Pareto para θ, $Pa(n-1,t)$, que é fechado sob produtos, pelo que a família Pareto é a conjugada natural do modelo amostral Uniforme acima definido. Assim, considerando $\theta \sim Pa(a,b)$ com $a,b>0$ tem-se $\theta|x \sim Pa(A,B)$ com $A=a+n$, $B=b+t$ com função densidade $h(\theta|x) = AB^A\theta^{-(A+1)}I_{[B,+\infty)}(\theta)$. A associada distribuição difusa é então definida pela distribuição imprópria $h(\theta) \propto \theta^{-1}I_{[0,+\infty)}(\theta)$, por vezes rotulada como $\theta \sim$ "$Pa(0,0)$", passando então a distribuição *a priori* própria $\theta \sim Pa(a,b)$ a ser visualizável como atualização da distribuição difusa através de uma pseudoamostra de dimensão a e máximo amostral b.

Cálculos simples permitem chegar às estimativas pontuais bayesianas moda $= B$, média $= BA/(A-1)$ (se $A>1$) e mediana $= B2^{1/A}$. A forma da densidade *a posteriori* de θ mostra que o intervalo de credibilidade HPD a $100\gamma\%$ é $(B,\bar{\theta})$ com limite superior $\bar{\theta} = B(1-\gamma)^{-1/A}$.

Para a predição de uma nova observação Y do modelo amostral supostamente independente das observações já feitas, a função densidade preditiva *a posteriori* apresenta os seguintes dois troços (sendo obviamente nula em $\mathbb{R}_-$):

$$p(y|x) = \begin{cases} \frac{A}{A+1}p_{U(]0,B])}(y), & \text{se } 0<y\leq B \\ \frac{1}{A+1}p_{Pa(A,B)}(y), & \text{se } y>B. \end{cases}$$

A forma desta densidade mostra que para $\gamma \geq \frac{A}{A+1} = P(0<Y\leq B|x)$ o intervalo de predição com a máxima densidade preditiva é definido por $\{y>B : p(y|x)\geq c_\gamma\} = (0,k_\gamma)$ com limite superior $k_\gamma = B\left[(A+1)(1-\gamma)\right]^{-1/A}$.

3.5 Modelo Normal (com média conhecida) ∧ Gama Inversa

Seja $x=(x_i,\ i=1,\ldots,n)$ uma concretização do vetor de variáveis aleatórias $X_i,\ i=1,\ldots,n \overset{iid|\sigma^2}{\sim} N(\mu_0,\sigma^2)$, com μ_0 conhecido. A correspondente função densidade, ao poder escrever-se como

$$f(x_1,\ldots,x_n|\sigma^2) \propto (\sigma^2)^{-n/2}e^{-\frac{\sum_i(x_i-\mu_0)^2}{2\sigma^2}},$$

é proporcional ao núcleo de uma distribuição Gama Inversa para σ^2, $GaI(\frac{n}{2}-1, \frac{1}{2}\sum_i(x_i-\mu_0)^2)$, que é fechado sob produtos. A conjugada natural deste

modelo amostral é então a família GaI pelo que $\sigma^2 \sim GaI(a,b) \Leftrightarrow \sigma^{2^{(-1)}} \sim Ga(a,b)$ implica que $\sigma^2|x \sim GaI(A,B)$ com $A = a + \frac{n}{2},\ B = b + \frac{1}{2}\sum_i (x_i - \mu_0)^2$. A distribuição *a priori* $GaI(a,b)$ é assim interpretável como a distribuição *a posteriori* resultante da atualização da distribuição vaga "$Ga(0,0)$" por uma pseudoamostra de tamanho $2a$ do correspondente modelo Normal de média zero com soma dos quadrados das observações igual a $2b$.

Inferências paramétricas sobre a dispersão ou precisão das observações amostrais obtêm-se sem grande dificuldade de distribuições GaI ou Ga (ou com elas relacionadas). Inferências preditivas sobre Y a tomar do mesmo modelo amostral independentemente de $X_i,\ i = 1,\ldots,n$ obtêm-se da mistura de $N(\mu_0, \sigma^2)$ pela distribuição *a posteriori* $\sigma^2|x \sim GaI(A,B)$, a qual define uma distribuição t-Student com $2A$ graus de liberdade, parâmetro de localização μ_0 e parâmetro de escala $\sqrt{B/A}$, simbolizada por $t_{(2A)}(\mu_0, B/A)$, com função densidade preditiva

$$\begin{aligned} p(y|x) &= \frac{B^A}{\sqrt{2\pi}\,\Gamma(A)} \int_0^{+\infty} (\sigma^2)^{-(A+3/2)} e^{-(1/\sigma^2)\left[B + \frac{(y-\mu_0)^2}{2}\right]} d\sigma^2 \\ &= \left[B(\frac{2A}{2}, \frac{1}{2})\right]^{-1} \left(\sqrt{2AB/A}\right)^{-1} \left[1 + \frac{(y-\mu_0)^2}{2AB/A}\right]^{-\frac{2A+1}{2}} \end{aligned}$$

3.6 Modelo Normal biparamétrico ∧ distribuições *a priori* marginais de Jeffreys

Denotando $x = (x_i,\ i = 1,\ldots,n)$ os dados relativos a uma amostra aleatória $(X_i,\ i = 1,\ldots,n)$ do modelo $N(\mu, \sigma^2)$, a função de verosimilhança

$$f(x_1,\ldots,x_n|\mu,\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2 - \frac{ks^2}{2\sigma^2}\right\},$$

onde $k = n - 1$ e $ks^2 = \sum_i (x_i - \bar{x})^2$, é o núcleo de uma distribuição conjunta Normal-Gama Inversa para (μ, σ^2) entendida como uma Normal para μ dado σ^2 e uma Gama Inversa para σ^2, que é fechado sob produtos. A família conjugada natural é então definida por funções densidade do tipo $h(\mu, \sigma^2|a,v,c,d) = h_{N(a,\sigma^2/v)}(\mu|\sigma^2)\, h_{GaI(c,d)}(\sigma^2)$.

A distribuição de Jeffreys sob independência *a priori* entre μ e σ^2, $h(\mu,\sigma^2) \propto \sigma^{-2}$ é um caso limite de uma distribuição imprópria N–GaI. A sua atualização

bayesiana pela verosimilhança indicada origina a função densidade de probabilidade

$$h(\mu,\sigma^2|x) \propto (\sigma^2)^{-1/2} e^{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2} \times (\sigma^2)^{-(\frac{n-1}{2}+1)} e^{-\frac{ks^2}{2\sigma^2}}$$

evidenciando que $\mu|\sigma^2, x \sim N(\bar{x}, \sigma^2/n)$ e $\sigma^2|x \sim GaI(\frac{k}{2}, \frac{ks^2}{2}) \Leftrightarrow \frac{ks^2}{2\sigma^2}|x \sim \chi^2_{(k)}$.

A distribuição marginal *a posteriori* de μ é então uma mistura de Normais por uma GaI, que já se sabe ser do tipo t-Student. Com efeito, por integração de σ^2, tendo em conta a função densidade GaI, resulta

$$h(\mu|x) = \left[B\left(\frac{k}{2},\frac{1}{2}\right)\right]^{-1} \left(\sqrt{ks^2/n}\right)^{-1} \left[1+\frac{(\mu-\bar{x})^2}{ks^2/n}\right]^{-\frac{k+1}{2}},$$

isto é, $\mu|x \sim t_{(k)}(\bar{x}, s^2/n) \Leftrightarrow \frac{\mu-\bar{x}}{s/\sqrt{n}}|x \sim t_{(k)}(0,1) \Leftrightarrow \frac{(\mu-\bar{x})^2}{s^2/n}|x \sim F_{(1,k)}$, onde a distribuição t-Student reduzida $t_{(k)}(0,1)$ é a conhecida t-Student da Estatística Clássica. Assim, $E(\mu|x) = \bar{x}$ (se $k > 1$) e $Var(\mu|x) = \frac{k}{k-2}\frac{s^2}{n}$ (se $k > 2$). A distribuição *a posteriori* de σ^2 condicional a μ pode então ser determinada como $\sigma^2|\mu, x \sim GaI(\frac{k+1}{2}, \frac{ks^2+n(\mu-\bar{x})^2}{2})$.

Inferências sobre cada um dos parâmetros de localização e escala do modelo amostral obtêm-se sem dificuldade de maior com base nas distribuições t-Student e Gama Inversa (ou χ^2 por transformação apropriada). Em termos de predição, considerando-se por exemplo uma amostra aleatória futura de tamanho m do modelo cuja média $\bar{Y}$ se pretende predizer, a correspondente distribuição preditiva *a posteriori* é a mistura das distribuições $\bar{Y}|\mu,\sigma^2 \sim N(\mu, \sigma^2/m)$ pela distribuição *a posteriori* conjunta $h(\mu,\sigma^2|x) = h(\mu|\sigma^2,x)h(\sigma^2|x)$. Tendo em conta a identidade algébrica da combinação linear das formas quadráticas, $\frac{1}{\sigma^2}\left[m(\mu-\bar{y})^2 + n(\mu-\bar{x})^2\right]$, essa distribuição reveste a forma t-Student, $\bar{Y}|x \sim t_{(k)}(\bar{x}, \frac{m+n}{mn}s^2)$, da qual se determinam facilmente resumos pontuais e intervalares.

3.7 Dois modelos Normais independentes ∧ distribuições marginais de Jeffreys

Sejam $x_j = (x_{ji},\ i = 1,\ldots,n_j)$, $j = 1,2$ concretizações de duas amostras aleatórias independentes de populações $N(\mu_j, \sigma_j^2)$ e use-se para os 4 parâmetros a distribuição *a priori* de Jeffreys comummente adotada $h(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) \propto (\sigma_1^2\sigma_2^2)^{-1}$ no respetivo espaço paramétrico conjunto.

Comparação de médias

Seguindo um raciocínio análogo ao da secção imediatamente anterior, conclui-se facilmente que (μ_1, σ_1^2) e (μ_2, σ_2^2) são *a posteriori* também independentes com as distribuições marginais univariadas

$$\mu_j|x_j \sim t_{(k_j)}(\bar{x}_j, s_j^2/n_j) \Leftrightarrow \nu_j = \frac{\mu_j - \bar{x}_j}{s_j/\sqrt{n_j}}|x_j \sim t_{(k_j)}$$

$$\sigma_j^2|x_j \sim GaI(\frac{k_j}{2}, \frac{k_j s_j^2}{2})$$

em que $k_j = n_j - 1$ e $k_j s_j^2 = \sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2$.

A redução de $\lambda = \mu_1 - \mu_2$, expressa por

$$\tau = \frac{\lambda - (\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \equiv \nu_1 \text{sen}\, u + \nu_2 \cos u,$$

em que $u = \arctan(\frac{s_1}{\sqrt{n_1}}/\frac{s_2}{\sqrt{n_2}})$, é distribuída *a posteriori* como uma combinação linear de distribuições t-Student independentes, conhecida como distribuição de Behrens-Fisher e parametrizada por k_1, k_2 e u [7]. A sua função densidade, que é simétrica mas não expressável em forma fechada, costuma ser na prática aproximada por uma distribuição t-Student devida a Patil (1964), concretamente $\tau|x_1, x_2 \sim BF(k_1, k_2, u) \underset{aprox}{\sim} t_{(b)}(0, a)$, onde

$$b = 4 + c_1^2/c_2, \;\; a = \sqrt{c_1(b-2)/b}$$

$$c_1 = \frac{k_1}{k_1 - 2}\text{sen}^2 u + \frac{k_2}{k_2 - 2}\cos^2 u$$

$$c_2 = \frac{k_1^2}{(k_1-2)^2(k_1-4)}\text{sen}^4 u + \frac{k_2^2}{(k_2-2)^2(k_2-4)}\cos^4 u \;.$$

Uma alternativa ao uso da aproximação de Patil consiste na geração de uma amostra da distribuição *a posteriori* de τ através de simulação a partir das distribuições *a posteriori* de ν_1 e ν_2, com base na qual se podem calcular empiricamente estimativas pontuais e intervalares e testar hipóteses pontuais sobre a diferença de médias.

[7] Note-se que esta dependência de u leva a que não haja dualidade entre as distribuições *a posteriori* e amostral de τ e, consequentemente, identidade numérica entre as inferências bayesiana e clássica sobre a diferença de médias, contrariamente ao que acontece em outras situações em que se usam distribuições *a priori* não informativas.

Comparação de variâncias

Tomando como parâmetro de interesse $\psi = \frac{\sigma_1^2}{\sigma_2^2}$, conclui-se das distribuições *a posteriori* Gama independentes de $\{1/\sigma_j^2\}$ que $\psi|x_1, x_2 \overset{d}{\equiv} \frac{s_1^2}{s_2^2} F_{(k_2,k_1)}$, o que permite realizar facilmente inferências básicas sobre ψ.

Comparação de médias de populações homocedásticas

Neste contexto admite-se $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$ e o uso da distribuição *a priori* de Jeffreys $h(\mu_1, \mu_2, \sigma^2) \propto \sigma^{-2}$, $\mu_1, \mu_2 \in I\!R$, $\sigma^2 > 0$. Tomando de novo como referência o material da secção anterior, rapidamente se conclui que

$$\lambda = \mu_1 - \mu_2|\sigma^2, x_1, x_2 \sim N(\bar{x}_1 - \bar{x}_2, \sigma^2(\tfrac{1}{n_1} + \tfrac{1}{n_2}));$$
$$\sigma^2|x_1, x_2 \sim GaI(\tfrac{k}{2}, \tfrac{ks^2}{2});$$

onde $k = n_1 + n_2 - 2$ e $s^2 = k^{-1} \sum_j (n_j - 1)s_j^2$ é a variância empírica combinada. Isto implica designadamente que

$$\lambda = \mu_1 - \mu_2|x_1, x_2 \sim t_{(k)}\left(\bar{x}_1 - \bar{x}_2, s^2(\frac{1}{n_1} + \frac{1}{n_2})\right) \Leftrightarrow \frac{\lambda - (\bar{x}_1 - \bar{x}_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}|x_1, x_2 \sim t_{(k)},$$

que é o resultado básico para o traçado das inferências de interesse sobre a comparação das duas populações Normais.

3.8 Dois modelos Binomiais independentes ∧ distribuições Beta

Sejam t_j, $j = 1, 2$ contagens observadas de T_j, $j = 1, 2|\theta_j \underset{ind}{\sim} Bi(m_j, \theta_j)$, com $\{m_j\}$ conhecidos e considere-se o modelo *a priori* θ_j, $j = 1, 2 \underset{ind}{\sim} Be(a_j, b_j)$. De acordo com o que se viu para o modelo Bi ∧ Be, resulta que

$$\theta_j|t_j, j = 1, 2 \underset{ind}{\sim} Be(A_j, B_j),\ A_j = a_j + t_j,\ B_j = b_j + m_j - t_j$$
$$\Leftrightarrow (B_j/A_j)\tfrac{\theta_j}{1-\theta_j}|t_j, j = 1, 2 \underset{ind}{\sim} F_{(2A_j, 2B_j)}$$
$$\Leftrightarrow \left[(1/2)\ln(B_j/A_j) + (1/2)\ln\tfrac{\theta_j}{1-\theta_j}\right]|t_j, j = 1, 2 \underset{ind}{\sim} Z_{(2A_j, 2B_j)}.$$

Testes unilaterais exatos de comparação de proporções

Para o teste de $H_0 : \theta_1 \leq \theta_2$ contra $H_1 : \theta_1 > \theta_2$ o uso das chances *a posteriori* ou do fator Bayes requer a avaliação das quantidades

$$P(H_0|t_1, t_2) = \int_0^1 h(\theta_1|t_1) \left[\int_{\theta_1}^1 h(\theta_2|t_2) d\theta_2 \right] d\theta_1$$

(simultaneamente com o seu análogo baseado em distribuições *a priori*). No caso de a_2 e b_2 serem inteiros, a relação entre as funções de distribuição Beta e Binomial, $F_{Be(A_2,B_2)}(\theta_1) = 1 - F_{Bi(A_2+B_2-1,\theta_1)}(A_2 - 1)$ permite que se escreva

$$P(H_0|t_1, t_2) = [B(A_1, B_1)]^{-1} \sum_{u=0}^{A_2-1} \binom{A_2 + B_2 - 1}{u} B(A_1 + u, B_1 + A_2 + B_2 - 1 - u),$$

com as funções beta calculáveis em termos de fatoriais se cumulativamente a_1 e b_1 também forem inteiros.

Testes de homogeneidade das Binomiais, $H_0 : \theta_1 = \theta_2$ contra $H_1 : \theta_1 \neq \theta_2$

Atendendo a que $H_0 : \pi = 0 \Leftrightarrow \ln \Delta = 0$ usando as transformações $\pi = \theta_1 - \theta_2$ e $\Delta = \frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)}$, o recurso a simulação a partir das distribuições *a posteriori* $Be(A_j, B_j)$ – e.g., via $\theta_j = \frac{\nu_{A_j}}{\nu_{A_j} + \nu_{B_j}}$, com $\nu_{K_j}, K = A, B \underset{ind}{\sim} Ga(K_j, 1)$ – de amostras de π ou de Δ (ou $\ln \Delta$) permite obter boas aproximações empíricas do nível de plausibilidade relativa *a posteriori* de H_0 ou de intervalos HPD. Obviamente que esta via é também aplicável a hipóteses unilaterais por meio do cálculo de apropriadas proporções com base nas amostras simuladas.

No caso de grandes valores observados de sucessos e insucessos, a utilização de aproximações assintóticas à distribuição Z de Fisher, por exemplo

$$Z_{(\nu_1,\nu_2)} \overset{aprox}{\sim} N\left[\frac{1}{2} \ln \frac{\nu_1^{-1} - 1}{\nu_2^{-1} - 1}, \ \frac{1}{2}(\nu_1^{-1} + \nu_2^{-1}) \right],$$

possibilita recorrer à distribuição *a posteriori* aproximada

$$\ln \Delta | t_1, t_2 \overset{aprox}{\sim} N\left[\ln \frac{(A_1 - 1/2)/(B_1 - 1/2)}{(A_2 - 1/2)/(B_2 - 1/2)}, \sum_{j=1,2} (A_j^{-1} + B_j^{-1}) \right]$$

para a construção de testes bayesianos unilaterais ou bilaterais das hipóteses em confronto.

3.9 Modelo Multinomial ∧ Dirichlet

Este modelo bayesiano é a versão multivariada do modelo Binomial ∧ Beta mas como é bem menos familiar do que este, opta-se por começar a descrever sucintamente as principais propriedades dele pela sua relevância no campo inferencial.

Seja $X = (X_1, \ldots, X_c)$ e $\theta = (\theta_1, \ldots, \theta_c)$ vetores aleatórios assumindo valores nos subespaços respetivamente $\mathcal{X} = \{x = (x_1, \ldots, x_c) : x_i \in I\!N_0, x_\bullet = \sum_{i=1}^{c} x_i \leq N\}$, onde N se supõe conhecido, e $\Theta = \{(\theta_1, \ldots, \theta_c) : \theta_i \in (0,1), \theta_\bullet = \sum_{i=1}^{c} \theta_i < 1\}$, o chamada simplex c-dimensional $\mathcal{S}_c$.

Distribuição Multinomial (c-variada) para X

Função de probabilidade de $X|\theta \sim M_c(N, \theta) : f(x|\theta) = \frac{N!}{\prod_{i=1}^{c+1} x_i!} \prod_{i=1}^{c+1} \theta_i^{x_i},\ x \in \mathcal{X}$ em que $x_{c+1} = N - x_\bullet$. Os seus dois primeiros momentos são definidos por (aplique-se, e.g., a técnica da função geradora de momentos (f.g.m.)):

$$\mu = E(X|\theta) = N\theta;\ \Sigma = Var(X|\theta) = N(D_\theta - \theta\theta'),$$

onde $D_\theta = \text{diag}(\theta_1, \ldots, \theta_c)$.

Denote-se por $C_k = \{j_{k-1} + 1, \ldots, j_k\},\ k = 1, \ldots, s+1$ as partes de uma partição do conjunto dos índices $\{1, 2, \ldots, c, c+1\}$ em $s+1$ subconjuntos em que $\#C_k = d_k,\ j_0 = 0$ e $j_{s+1} = c+1$. Em correspondência, considerem-se os desdobramentos das componentes de X definidos por

$$M_k = \sum_{i \in C_k} X_i, \quad k = 1, \ldots, s+1; \qquad X^{(k)} = (X_i, i \in C_k), \quad k = 1, \ldots, s+1.$$

Então (recorra-se à via da f.g.m. e da definição de distribuição condicional)

$$M = (M_1, \ldots, M_s)|\theta \sim M_s(N, \alpha),\ \alpha = (\alpha_1, \ldots, \alpha_s),\ \alpha_k = \textstyle\sum_{i \in C_k} \theta_i$$
$$X^{(k)}|M, \theta,\ k = 1, \ldots, s+1 \underset{ind}{\sim} M_{d_k - 1}(M_k, \pi_k), \quad \pi_k = (\theta_i/\alpha_k,\ i \in C_k - j_k).$$

Estas transformações evidenciam que distribuições marginais e condicionais apropriadas de componentes de uma distribuição Multinomial são também Multinomiais (Binomiais, no caso univariado). A sua relevância é cabalmente esclarecida no contexto de tabelas de contingência. Por exemplo,

se X representa o vetor das frequências de uma tabela de contingência bidimensional, as respetivas frequências da margem das linhas (ou das colunas) são expressas adequadamente por M e as frequências das linhas (ou das colunas) condicionadas aos totais marginais são caraterizadas pela distribuição condicional indicada Produto de Multinomiais.

Distribuição Dirichlet (c-variada) para θ

Função densidade de probabilidade de $\theta|a \sim D_c(a) : h(\theta|a) = [B(a)]^{-1} \times \prod_{i=1}^{c+1} \theta_i^{a_i-1}$, $\theta \in \Theta = \mathcal{S}_c$ em que $a = (a_1, \ldots, a_c, a_{c+1}) \in I\!R_+^c$, $\theta_{c+1=1-\theta_\bullet}$ e $B(a) = \frac{\prod_{i=1}^{c+1} \Gamma(a_i)}{\Gamma(a_\bullet)}$ é a função beta multivariada.

A distribuição Dirichlet pode ser vantajosamente definida nos planos teórico e computacional a partir de distribuições Gama independentes através da transformação $\theta_i = \frac{\nu_i}{\sum_{j=1}^{c+1} \nu_j}$, $i = 1, \ldots, c$, com as variáveis aleatórias ν_i, $i = 1, \ldots, c+1 \underset{ind}{\sim} Ga(a_i, 1)$.

Os momentos mistos são pela sua definição dados por $E\left[\prod_{i=1}^{c+1} \theta_i^{r_i}|a\right] = \frac{B(a+r)}{B(a)}$, onde $r = (r_1, \ldots, r_c, r_{c+1})$, de onde se obtém

$$E(\theta_i|a) \equiv E_i = \frac{a_i}{a_\bullet}; \; Var(\theta_i|a) \equiv V_i = \frac{E_i(1-E_i)}{a_\bullet+1};$$
$$cov(\theta_i, \theta_j|a) \equiv V_{ij} = -\frac{E_i E_j}{a_\bullet+1}, \; i \neq j.$$

Como $(\alpha, \pi_1, \ldots, \pi_{s+1})$ é uma reparametrização de θ prova-se (use-se, e.g., a representação por distribuições Gama) que

$$\alpha|a \sim D_s(\textstyle\sum_{i \in C_k} a_i, \; k = 1, \ldots, s+1)$$
$$\pi_k|a, \; k = 1, \ldots, s+1 \underset{ind}{\sim} D_{d_k-1}(a_i, i \in C_k).$$

Distribuição Multinomial-Dirichlet (c-variada) para X

Esta distribuição, também conhecida por distribuição de Pólya, traduz a mistura de uma Multinomial por uma Dirichlet, $X|a \sim MD_c(N, a)$, cuja função de probabilidade é dada por

$$p(x|a) = \frac{N!}{\prod_{i=1}^{c+1} x_i!} \frac{B(a_1 + x_1, \ldots, a_{c+1} + x_{c+1})}{B(a_1, \ldots, a_{c+1})}, \; x \in \mathcal{X}$$

e cujos dois primeiros momentos são expressáveis (use-se, e.g., as propriedades da esperança condicional) por

$$E(X|a) = N\frac{a}{a_.};\ V(X|a) = \frac{a_. + N}{a_.(a_. + 1)}N(D_a - \frac{aa'}{a_.}).$$

Aplicando os desdobramentos anteriores de X, facilmente se conclui que

$$M = (M_1, \ldots, M_s)|a \sim MD_s(N; \textstyle\sum_{i\in C_k} a_i,\ k = 1, \ldots, s+1)$$
$$X^{(k)}|M, a,\ k = 1, \ldots, s+1 \underset{ind}{\sim} MD_{d_k-1}(M_k; a_i, i \in C_k).$$

Aplicação inferencial

Sendo $x = (x_1, \ldots, x_c)$ uma concretização de um vetor aleatório $X|\theta \sim M_c(N, \theta)$, como $f(x|\theta)$ é proporcional ao núcleo de uma distribuição $D_c(x_i + 1, i = 1, \ldots, c+1)$ que é fechado sob produtos, conclui-se que a família Dirichlet é a conjugada natural de uma amostragem Multinomial. Assim, se $\theta|a \sim D_c(a)$ então $\theta|a, x \sim D_c(A),\ A = (A_i = a_i + x_i,\ i = 1, \ldots, c+1)$.

Estimativas bayesianas de $(\theta_i,\ i = 1, \ldots, c+1)$ podem ser traduzidas em particular pelas componentes da moda *a posteriori* $(A - 1_{c+1})/A_.$ (se $A_i > 1, \forall i$), onde 1_{c+1} representa um vetor de $c+1$ elementos iguais a 1, ou da média *a posteriori* $A/A_.$. Note-se como esta é uma média ponderada do vetor média *a priori*, $a/a_.$ e do vetor das proporções amostrais $p = (x_i/N,\ i = 1, \ldots, c+1)$.

Em análise de tabelas de contingência o interesse inferencial diz muitas vezes respeito a estruturas de independência (ou a outros modelos log-lineares) nas quais desempenham um papel crucial funções paramétricas de tipo $\sum_i b_i \ln \theta_i$ com $\sum_i b_i = 0$ (*vide e.g.* Paulino e Singer, 2006). Quando as componentes de A são grandes, pode-se invocar a normalidade aproximada da sua distribuição *a posteriori* e a decorrente distribuição Qui-quadrado para apropriadas formas quadráticas, permitindo testar aquelas estruturas (para aprofundamento *vide e.g.* Cap. 6 de Paulino, Amaral Turkman e Murteira, 2003). No caso especial de uma tabela 2×2 a aplicação deste procedimento à hipótese de independência, que mais não é do que a de homogeneidade de duas Binomiais, conduz à via mencionada no fim da secção anterior.

Se o objetivo for o de predizer um vetor Y tal que $Y|m, \theta \sim M_c(m, \theta)$, a correspondente distribuição preditiva *a posteriori* é Multinomial-Dirichlet, $Y|m, x \sim MD_c(m, A)$, cujo resumo através dos seus dois primeiros momentos é obtenível das fórmulas expostas atrás.

3.10 Inferência sobre populações finitas

Considere-se uma população finita de tamanho conhecido N particionada em $c \leq N$ grupos de tamanhos desconhecidos $N_i,\ i = 1, \ldots, c$, tal que $\sum_{i=i}^{c} N_i = N$, da qual se selecionou aleatoriamente (sem reposição) uma amostra $\mathcal{S}$ de $n \leq N$ unidades com o objetivo de traçar inferências sobre o vetor de totais populacionais dos grupos $\theta = (N_1, \ldots, N_c)$. Sejam $n_i,\ i = 1, \ldots, c$ as frequências observadas dos grupos $(\sum_{i=1}^{c} n_i = n)$ que agrupamos em $x = (n_1, \ldots, n_c)$ que, pelo exposto, é uma observação da distribuição Hipergeométrica multivariada $X|N, n, \theta \sim Hpg_{c-1}(\theta, n)$ (usa-se, aqui e em seguida, por conveniência uma redundância notacional na definição de vetores aleatórios como em X.

Denotando por U_k o vetor indicador do grupo a que pertence a unidade k, cujos valores possíveis são os vetores da base ortonormal padrão de $\mathbb{R}^c$, o alvo inferencial é expressável por

$$\theta = \sum_{k=1}^{N} U_k = \sum_{k \in \mathcal{S}} U_k + \sum_{k \notin \mathcal{S}} U_k \equiv X + (\theta - X),$$

evidenciando particularmente que *a posteriori* apenas $\theta - X$ é desconhecido.

Admita-se uma distribuição *a priori* numa estrutura hierárquica definida por $U_1, \ldots, U_N \underset{iid}{\sim} M_{c-1}(1, \phi)$ dado algum parâmetro subjacente $\phi = (\phi_j,\ j = 1, \ldots, c)$, com $\sum_j \phi_j = 1$, a que se atribui num 2º nível a distribuição $\phi|a \sim D_{c-1}(a), a = (a_1, \ldots, a_c) \in \mathbb{R}_+^c$.

Em termos do 1º nível da hierarquia, tem-se assim que $\theta|\phi \sim M_{c-1}(N, \phi)$ e que X e $\theta - X$, por definição, são *a priori* e condicionalmente a ϕ independentemente distribuídos segundo leis do mesmo tipo, $X|n, \phi \sim M_{c-1}(n, \phi)$ e $\theta - X|n, \phi \sim M_{c-1}(N - n, \phi)$. Observe-se ainda que a distribuição amostral Hipergeométrica de X pode ser encarada como

$$f(x|n, \theta) = \frac{\prod_{j=1}^{c} \binom{N_j}{x_j}}{\binom{N}{n}} = \frac{f(x|n, \phi)\, h(\theta - x|n, \phi)}{h(\theta|\phi)} = \frac{f(x, \theta|n, \phi)}{h(\theta|\phi)} = f(x|n, \theta, \phi).$$

Usando a informação do 2º nível, pode-se identificar como sendo do tipo Multinomial-Dirichlet as seguintes distribuições marginais (ou preditivas *a priori*)

$$\theta|a \sim MD_{c-1}(N, a);\ X|n, a \sim MD_{c-1}(n, a);\ \theta - X|n, a \sim MD_{c-1}(N - n, a)$$

e que a atualização por x de ϕ é tal que $\phi|x \sim D_{c-1}(a+x)$. Por outro lado, como $\theta - x|x, \phi \stackrel{d}{=} \theta - x|\phi \sim M_{c-1}(N-n, \phi)$, resulta que $\theta - x|x \sim MD_{c-1}(N-n, a+x)$.

Em suma, a distribuição *a posteriori* de $\theta - x$ sob uma amostragem Hipergeométrica é do mesmo tipo (Multinomial-Dirichlet) da respetiva distribuição *a priori* e a distribuição *a posteriori* do vetor de totais θ resulta daquela por uma translação de x, da qual se obtém trivialmente a do vetor de proporções populacionais θ/N (Basu e Pereira, 1982).

Capítulo 4

Inferências por Métodos de Monte Carlo

A maioria dos problemas estatísticos abrange modelos complexos que tornam frequentemente irrealizável o cálculo por via analítica (ou mesmo numérica) de quantidades de interesse expressas por integrais. Neste quadro os métodos de Monte Carlo clássicos surgem muitas vezes como uma alternativa apropriada, ao basearem a determinação de inferências relevantes em cálculos envolvendo amostras simuladas de pertinentes distribuições de probabilidade, obtidas a partir de geradores de números (pseudo)aleatórios (valores da distribuição uniforme $U(0,1)$). Os procedimentos de simulação estocástica têm sido objeto de descrição em muita da literatura[8] e são atualmente executáveis por muito do software estatístico e matemático disponível.

Neste capítulo descreve-se a ideia geral subjacente aos métodos de Monte Carlo tradicionais nas versões simples e com a assim chamada amostragem de importância (*Importance Sampling*), bem como algumas das suas especializações (ou variantes) atinentes a inferências bayesianas potencialmente relevantes para o problema estatístico em mão.

[8] Destacam-se os livros de Devroye (1986) - edição *web* munida de errata e disponibilizada pelo autor em 2003 -, Ripley (1987) e Gentle (2004).

4.1 Monte Carlo simples

Considere-se o problema de aproximar um integral da forma

$$\int g(\theta)h(\theta|x)d\theta = E[g(\theta)|x], \tag{4.1}$$

onde θ e x podem ser vetores, cuja existência se admite. Muitas quantidades *a posteriori* de interesse são expressáveis por (4.1) para algum tipo de função $g(\theta)$ integrável. É o caso dos momentos *a posteriori* de componentes de θ, probabilidades *a posteriori* de subconjuntos do espaço paramétrico e densidades preditivas *a posteriori*, em que $g(\theta)$ é, respetivamente, representada por θ_i (para o caso da média da i-ésima componente de θ), $I_A(\theta)$ para $A \subset \Theta$ e $f(y \mid \theta)$ para y fixado.

Outras quantidades de interesse podem ainda ser expressas através de integrais apropriados como é o caso das constantes normalizadoras das distribuições *a posteriori*, densidades marginais *a posteriori*, fatores de Bayes e probabilidades *a posteriori* de modelos.

Se se puder simular uma amostra aleatória $\theta_1, \ldots, \theta_n$ da densidade *a posteriori* $h(\theta \mid x)$, o método de Monte Carlo simples aproxima o integral (4.1) pela média empírica

$$\hat{E}\,[g(\theta) \mid x] = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \tag{4.2}$$

a qual, pela Lei Forte dos Grandes Números, converge quase certamente para $E\,[g(\theta) \mid x]$. A precisão deste estimador pode ser medida pelo erro padrão (estimado) de Monte Carlo dado por

$$\frac{1}{\sqrt{n(n-1)}}\left\{\sum_{i=1}^{n}\left[g(\theta_i) - \frac{1}{n}\sum_{i=1}^{n} g(\theta_i)\right]^2\right\}^{1/2}, \tag{4.3}$$

quando a quantidade $E\{[g(\theta)]^2|x]$ é finita.

O integral de Riemann de interesse (4.1) pode ser representado de infinitos modos por alteração consistente do terno (Θ, g, h) (veja-se Ripley, 1987). Os estimadores Monte Carlo associados com as diversas representações apresentam precisões variáveis, com implicações no esforço computacional (modo mais ou menos fácil de simulação e dimensão maior ou menor da amostra) requerido para obtenção de estimativas fiáveis. Isto sugere a opção por vias

de maior eficiência de modo a obter estimadores altamente precisos com um número relativamente baixo de valores simulados [9].

Em suma, se se conseguir simular amostras da distribuição *a posteriori* $h(\theta \mid x)$, a aplicação do método de Monte Carlo simples para resolver integrais do tipo (4.1) é então trivial. Sublinhe-se, desde já, que a realização de inferências a partir da amostra simulada simplifica consideravelmente operações que analiticamente podem ser tremendamente complicadas. É o caso das reparametrizações e marginalizações que são simplesmente tratadas através das correspondentes transformações da amostra simulada e da seleção dirigida das respetivas componentes dos vetores simulados. As subsecções que se seguem especializam o método de Monte Carlo simples na avaliação de probabilidades *a posteriori*, densidades *a posteriori* marginais, intervalos de credibilidade e quantidades associadas à distribuição preditiva *a posteriori*.

4.1.1 Probabilidades *a posteriori*

Quando $g(\theta)$ é a função indicadora de algum subconjunto A do espaço paramétrico, a aproximação de Monte Carlo (4.2) representa a proporção de valores amostrais incluídos em A. Uma exemplificação concreta deste caso reporta-se ao cálculo da probabilidade *a posteriori* do menor intervalo HPD contendo um valor fixado $\theta_0 \in \mathbb{R}$,

$$P(\theta_0) = P_{h(\theta|x)}\left(\{\theta : h(\theta \mid x) \geq h(\theta_0 \mid x)\}\right),$$

referido no Capítulo 1 como meio de construir testes de significância bayesianos para hipóteses $H_0 : \theta = \theta_0$. A determinação deste nível de plausibilidade relativa *a posteriori* dispensa o conhecimento da constante normalizadora da densidade univariada $h(\theta \mid x)$ e a correspondente estimativa de Monte Carlo pode ser assim expressa por

$$\hat{P}(\theta_0) = \frac{1}{n} \#\left\{\theta_i,\ 1 \leq i \leq n : L(\theta_i \mid x)\, h(\theta_i) \geq L(\theta_0 \mid x)\, h(\theta_0)\right\}. \tag{4.4}$$

4.1.2 Intervalos de credibilidade

Considere-se agora que $(\theta_i,\ 1 \leq i \leq n)$ é uma amostra aleatória da densidade *a posteriori* univariada $h(\theta \mid x)$, com função de distribuição $H(\theta \mid x)$, que

[9] Para conhecimento de técnicas de redução de variância na estimação por Monte Carlo veja-se, *e.g.*, Rubinstein (1981) e Robert e Casella (2004).

pretende ser resumida por um intervalo de credibilidade γ. A determinação exata deste exige o conhecimento completo da distribuição *a posteriori* para obtenção dos correspondentes quantis. No caso de desconhecimento da constante normalizadora, pode aproveitar-se a referida amostra para obter uma aproximação Monte Carlo do intervalo de credibilidade através dos respetivos quantis empíricos.

Uma aproximação Monte Carlo de $R_c(\gamma)$ é obtida ordenando a amostra aleatória e usando os quantis empíricos. Especificamente, representando agora $(\theta_{(i)}, 1 \leq i \leq n)$ a amostra ordenada, a estimativa Monte Carlo de $R_c(\gamma)$ é definida por

$$\hat{R}_c(\gamma) = \left(\theta_{([n\frac{1-\gamma}{2}])}, \theta_{([n\frac{1+\gamma}{2}])}\right), \tag{4.5}$$

onde $[n\alpha]$ denota a parte inteira de $n\alpha$.

O melhor resumo intervalar de uma distribuição unimodal quando assimétrica é o intervalo HPD $R_0(\gamma) = \{\theta : h(\theta \mid x) \geq k_\gamma\}$ onde k_γ é a maior constante para a qual a probabilidade *a posteriori* de $R_0(\gamma)$ é no mínimo γ. Pela sua definição, este intervalo é mais difícil de determinar do que os intervalos de caudas com áreas preestabelecidas, mesmo que se disponha de formas fechadas para as funções densidade e de distribuição *a posteriori* de θ.

Chen e Shao (1999) propõem um procedimento de Monte Carlo para a aproximação de $R_0(\gamma)$ extremamente simples de aplicar. Com base na amostra ordenada, denotada por $(\theta_{(i)}, 1 \leq i \leq n)$, determinam-se os intervalos de credibilidade γ

$$\hat{R}_i(\gamma) = \left(\theta_{(i)}, \theta_{(i+[n\gamma])}\right),\ i = 1, \ldots, n - [n\gamma],$$

em que $[n\gamma]$ denota a parte inteira de $n\gamma$. Tendo em conta a propriedade de amplitude mínima dos intervalos HPD, a aproximação Monte Carlo de $R_0(\gamma)$ segundo o método de Chen-Shao é definido por $\hat{R}_0(\gamma) = R_{i_0}(\gamma)$ tal que $\theta_{(i_0+[n\gamma])} - \theta_{(i_0)} = \min\left[\theta_{(i+[n\gamma])} - \theta_{(i)}\right],\ 1 \leq i \leq n - [n\gamma]$[10].

Observe-se que este método é de aplicabilidade linear quando se pretendem determinar intervalos HPD para funções paramétricas $\psi(\theta)$ – recorde-se que o carácter HPD não é invariante sob transformações não lineares. Basta aplicá-lo à amostra transformada $(\psi(\theta_i),\ 1 \leq i \leq n)$.

[10]Sobre a validade assintótica da aproximação $\hat{R}_0(\gamma)$ veja-se Chen *et al.*, 2000, Cap. 7.

4.1.3 Densidades *a posteriori* marginais

Se $\theta = (\theta_1, \ldots, \theta_k) \in I\!R^k$, $k > 1$ e o objetivo é avaliar densidades marginais *a posteriori* com base numa amostra aleatória $\theta_{(i)} = (\theta_{(i)1}, \ldots, \theta_{(i)k})$, $1 \leq i \leq n$ de $h(\theta \mid x)$, é possível aplicar mais do que um método.

Quando o interesse está na avaliação de densidades marginais univariadas, diga-se $h(\theta_j \mid x)$, o método mais simples consiste na seleção das j-ésimas componentes da amostra multivariada e no traçado de um histograma, com base na decorrente amostra univariada $(\theta_{(1)j}, \ldots, \theta_{(n)j})$, ao qual se ajusta de forma expedita uma curva por algum simples procedimento de suavização. Um procedimento de alisamento não paramétrico mais sofisticado para a avaliação da densidade marginal $h(\theta^{(m)} \mid x)$, em que $\theta^{(m)} = (\theta_1, \ldots, \theta_m) \in I\!R^m$ para $m = 1, \ldots, k-1$ fixado, é o chamado método do núcleo (*kernel method*) cuja descrição pode ser encontrada em livros de inferência não paramétrica, por exemplo Silverman (1986).

Para apresentação dum outro método, que se vai basear na ideia de condicionamento, considere-se momentaneamente $k = 2$ e seja Θ o suporte da densidade *a posteriori* de $\theta = (\theta_1, \theta_2)$, $h(\theta_1, \theta_2 \mid x)$. Denote-se o subconjunto de Θ que representa o suporte de $h(\theta_1, \theta_2 \mid x)$ para θ_1 fixado por $\Theta_{-1}(\theta_1) = \{\theta_2 : (\theta_1, \theta_2) \in \Theta\}$ e o suporte da densidade condicional $h(\theta_1 \mid \theta_2, x)$ por $\Theta_1(\theta_2) = \{\theta_1 : (\theta_1, \theta_2) \in \Theta\}$.

Fixado um valor θ_{1*} de θ_1, tem-se (pressupondo a validade do teorema de Fubini)

$$\begin{aligned} h(\theta_{1*} \mid x) &= \int_{\Theta_{-1}(\theta_{1*})} h(\theta_{1*} \mid \theta_2, x) h(\theta_2 \mid x)\, d\theta_2 \\ &= \int_{\Theta_{-1}(\theta_{1*})} h(\theta_{1*} \mid \theta_2, x) \int_{\Theta_1(\theta_2)} h(\theta_1, \theta_2 \mid x)\, d\theta_1 d\theta_2 \\ &= \int_{\Theta} h(\theta_{1*} \mid \theta_2, x) h(\theta \mid x)\, d\theta, \end{aligned} \tag{4.6}$$

evidenciando como as ordenadas da densidade marginal *a posteriori* de θ_1 são interpretáveis como valores esperados *a posteriori* das correspondentes ordenadas da densidade *a posteriori* condicional em θ_2.

Generalizando este raciocínio para o particionamento $\theta = \left(\theta^{(m)}, \theta^{(-m)}\right)$, com $\theta^{(-m)} = (\theta_{m+1}, \ldots, \theta_k)$, obtém-se

$$h(\theta_*^{(m)} \mid x) = \int_{\Theta} h(\theta_*^{(m)} \mid \theta^{(-m)}, x) h(\theta \mid x)\, d\theta. \tag{4.7}$$

Esta expressão implica que a densidade marginal *a posteriori* de $\theta^{(m)} =$

$(\theta_1, \ldots, \theta_m)$ pode ser aproximada, segundo o método de Monte Carlo aplicado à amostra aleatória de $h(\theta \mid x)$, $\theta_{(i)} = (\theta_{(i)}^{(m)}, \theta_{(i)}^{(-m)})$ com $\theta_{(i)}^{(m)} = \left(\theta_{(i)1}, \ldots, \theta_{(i)m}\right)$ e $\theta_{(i)}^{(-m)} = \left(\theta_{(i)m+1}, \ldots, \theta_{(i)k}\right)$, $i = 1, \ldots, n$, por

$$\hat{h}\left(\theta_*^{(m)}\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(\theta_*^{(m)} \mid \theta_{(i)}^{(-m)}, x\right). \tag{4.8}$$

Esta estimativa condicional da densidade $h(\theta^{(m)} \mid x)$, proposta por Gelfand e Smith (1990) – veja-se também Gelfand, Smith e Lee (1992) –, não usa a parte $\theta_{(i)}^{(m)}$, $i = 1, \ldots, n$ dos valores simulados em que se concentra o método do núcleo, mas exige o conhecimento completo da densidade *a posteriori* condicional de $\theta^{(m)}$ dado $\theta^{(-m)}$.

Uma vez verificado esse pressuposto, a estimativa (4.8), ao tirar partido do conhecimento de parte da estrutura do modelo consubstanciada na referida distribuição condicional, revela-se mais eficiente do que a estimativa obtida pelo método *kernel*, como foi evidenciado por Gelfand e Smith (1990). Algo análogo sucede com a estimativa da média *a posteriori* de $\theta^{(m)}$ obtida, tendo em conta (4.8), por

$$\hat{\theta}^{(m)} = \frac{1}{n}\sum_{i=1}^{n} E\left(\theta^{(m)} \mid \theta_{(i)}^{(-m)}, x\right). \tag{4.9}$$

Esta estimativa, requerendo o conhecimento do valor esperado condicional indicado, é mais precisa do que a estimativa clássica de Monte Carlo obtida de uma amostra gerada da distribuição marginal de $\theta^{(m)}$, devido a propriedades da esperança condicional[11].

4.1.4 Quantidades preditivas

Atendendo a que as ordenadas da densidade preditiva *a posteriori* de Y são o valor esperado $p(y \mid x) = E_{\theta|x}\left[f(y \mid \theta, x)\right]$, facilmente se obtém a respetiva

[11] Este é um argumento análogo ao usado pelo teorema de Rao-Blackwell, o qual estabelece que o valor esperado de um estimador condicional numa estatística suficiente é um estimador com o mesmo viés mas mais eficiente do que o primeiro. Ainda que o contexto aqui não seja o mesmo, o rótulo *Rao-Blackwellization* colado ao processo de condicionamento passou a ser usado na literatura para os estimadores condicionais (4.9) e (4.8) - daí o neologismo *Rao-Blackwellização* que livremente se tem proposto.

aproximação de Monte Carlo

$$\hat{p}(y \mid x) = \frac{1}{n} \sum_{i=1}^{n} f(y \mid \theta_i, x) \tag{4.10}$$

com base nos valores i.i.d. simulados de $h(\theta \mid x)$.

Para a estimação Monte Carlo de quantidades associadas com a distribuição preditiva $p(y \mid x)$ é necessário obter-se uma amostra aleatória desta distribuição. Isto é possível através do chamado método de composição (Tanner, 1996, Sec. 3.3) caso se saiba amostrar da distribuição amostral de y, obtendo-se então a amostra $(y_1, \ldots, y_n)$ de $p(y \mid x)$ do seguinte modo:

- Retire-se uma amostra de valores i.i.d. de $h(\theta \mid x)$, $(\theta_1, \ldots, \theta_n)$;
- Para cada i, gere-se y_i de $f(y \mid \theta_i, x)$, $i = 1, \ldots, n$.

Com base nesta amostra podem calcular-se facilmente aproximações de vários resumos da distribuição preditiva. Por exemplo, estimativas da predição média e do intervalo de predição HPD para a observação futura $y \in \mathbb{R}$ obtêm-se dela pela mesma forma como a média *a posteriori* e os intervalos de credibilidade HPD para θ são estimados da amostra da distribuição *a posteriori* de θ, como se indicou atrás.

4.2 Monte Carlo com amostragem de importância

Embora haja diversos métodos para simular amostras de várias distribuições (veja-se, *e.g.*, Ripley, 1987), geralmente não é possível obter uma amostra i.i.d. directamente da distribuição *a posteriori* $h(\theta|x)$ e, assim, há necessidade de encontrar estratégias alternativas. Por exemplo, uma das estratégias possíveis é a de simular de uma distribuição "semelhante" à distribuição *a posteriori*. Insere-se nesta linha a denominada amostragem de importância que se passa a descrever.

Seja $p(\theta)$ uma função densidade cujo suporte (diga-se Θ_p) inclua o de $h(\theta|x) = cf(x|\theta)h(\theta)$ e que se propõe como instrumento de amostragem. O valor esperado da função $g(\theta)$ segundo a distribuição *a posteriori* de θ, tomada como quantidade de interesse, pode exprimir-se em ordem a esta distribuição

$p(\theta)$ como o valor esperado da função original g ajustada pelo fator multiplicativo $h(\theta|x)/p(\theta)$, o qual é sempre finito pela condição imposta sobre o suporte da distribuição instrumental, $p(\theta)$. À luz do que se disse na secção anterior, a proposta de simular de $p(\theta)$ em vez de $h(\theta|x)$ conduz a redefinir a quantidade de interesse através do terno $(\Theta_p, \frac{gh}{p}, p)$.

Por outro lado, esta nova representação da quantidade de interesse exige apenas que a distribuição *a posteriori* seja conhecida a menos da constante de proporcionalidade c, sendo esta observação também aplicável à distribuição proposta para amostragem. Com efeito,

$$\begin{aligned} \int g(\theta)h(\theta|x)d\theta &= \frac{\int g(\theta)f(x|\theta)h(\theta)d\theta}{\int f(x|\theta)h(\theta)d\theta} = \frac{\int g(\theta)\frac{f(x|\theta)h(\theta)}{p(\theta)}p(\theta)d\theta}{\int \frac{f(x|\theta)h(\theta)}{p(\theta)}p(\theta)d\theta} \\ &\equiv \frac{\int g(\theta)w(\theta)p(\theta)d\theta}{\int w(\theta)p(\theta)d\theta} \end{aligned} \tag{4.11}$$

A densidade instrumental passou a ser conhecida por função de importância, possivelmente por permitir desbloquear o processo de simulação e cobrir melhor a região de importância para avaliação do integral em causa, passando então a obtenção da amostra simulada de tal função a ficar rotulada de amostragem de importância.

Sendo então $(\theta_1, \ldots, \theta_n)$ uma amostra de $p(\theta)$, pode-se aplicar o método de Monte Carlo, obtendo-se então como aproximação de $E[g(\theta) \mid x]$

$$\hat{E}[g(\theta) \mid x] = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i g(\theta_i), \tag{4.12}$$

onde $w_i = f(x \mid \theta_i)h(\theta_i)/p(\theta_i)$. A forma deste estimador mostra que a amostragem de importância pode ser vista como uma amostragem ponderada com os chamados pesos de importância w_i atribuídos a $g(\theta_i)$.

Nas condições admitidas, *i.e.* o suporte de $p(\theta)$ incluir o de $h(\theta|x)$ e o integral $\int g(\theta)h(\theta|x)d\theta$ existir e ser finito, Geweke (1989) mostra quando os θ_i são uma amostra i.i.d. de $p(\theta)$ que

$$\frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i g(\theta_i) \quad \to \quad \int g(\theta)h(\theta|x)d\theta \qquad q.c.,$$

com um erro padrão de Monte Carlo estimado por

$$\frac{1}{\sum_{j=1}^n w_j} \left[\sum_{i=1}^n \left\{ g(\theta_i) - \frac{1}{\sum_{j=1}^n w_j} \sum_{i=1}^n w_i g(\theta_i) \right\}^2 w_i^2 \right]^{1/2},$$

sob a finitude da variância do estimador de Monte Carlo, ou seja, do valor esperado *a posteriori* do produto de $[g(\theta)^2$ pelo rácio de importância $h(\theta|x)/p(\theta)$ (que é o valor esperado segundo p do quadrado de $g(\theta)h(\theta|x)/p(\theta)$).

A taxa de convergência do estimador Monte Carlo com amostragem de importância depende da relação entre a distribuição instrumental e a distribuição alvo ($h(\theta|x)$). Note-se que este estimador confere maior peso aos valores θ satisfazendo $p(\theta) < h(\theta|x)$ e menos peso àqueles para os quais se verifica o oposto. Se o rácio de importância for ilimitado, como acontece quando as caudas de p são mais leves que as de $h(\cdot|x)$, os pesos oscilarão bastante conferindo importância exagerada a poucos valores simulados, por se situarem em zonas (caudas) pouco plausíveis. Dependendo da função g, a variância deste estimador poderá mesmo ser infinita, implicando que venha a ter um pior desempenho do que o estimador obtido diretamente da distribuição alvo, quando possível.

Em suma, o método de Monte Carlo com amostragem de importância poderá constituir ou não uma bem-sucedida técnica de redução de variância dependendo da distribuição instrumental que for escolhida (*vide* Robert e Casella, 2004, para mais detalhes). "Boas" propriedades da função de importância são: (i) simplicidade na geração de números pseudoaleatórios; (ii) ter caudas mais pesadas do que $h(\cdot \mid x)$; (iii) ser uma boa aproximação para $h(\cdot \mid x)$. Shaw (1988) desenvolveu uma classe de distribuições univariadas que são adequadas para funcionar como funções de importância (veja-se também Smith, 1991). No caso multiparamétrico, utiliza-se com frequência distribuições normais multivariadas ou distribuições Student multivariadas[12].

Na mesma linha do disposto na secção anterior, passa-se agora a descrever a aplicação deste método (ou de alguma variante) no traçado de inferências associadas com intervalos de credibilidade, fatores de Bayes e densidades *a posteriori* marginais[13].

[12] Existe um pacote BAYESPACK que fornece uma interface R para sub-rotinas em Fortran que permitem fazer integração numérica e amostragem de importância, baseada em métodos descritos em Genz e Kass (1997). Este pacote pode ser obtido via *http://cran.r-project.org/src/contrib/Archive/bayespack*.

[13] Aspetos adicionais podem ser vistos nos livros Paulino *et al.* (2003) e Chen *et al.* (2000).

4.2.1 Intervalos de credibilidade

Para conferir um maior cunho de generalidade considera-se agora que o parâmetro do modelo amostral é particionado em $\theta = (\gamma, \phi)$, onde γ é o parâmetro de interesse unidimensional e ϕ é um vetor de parâmetros perturbadores (na parte correspondente da Secção 4.1, ϕ não existe pura e simplesmente, pelo que $\theta = \gamma$).

Sendo (γ_i, ϕ_i), $1 \le i \le n$, uma amostra aleatória de uma densidade $p(\gamma, \phi)$, que não é necessariamente a densidade conjunta *a posteriori* $h(\gamma, \phi \mid x)$, o valor da função de distribuição marginal *a posteriori* de γ no ponto γ_*

$$\begin{aligned} H(\gamma_* \mid x) &= E\left[I_{(-\infty,\gamma_*)}(\gamma) \mid x\right] \\ &= \frac{\int I_{(-\infty,\gamma_*)}(\gamma) \frac{L(\gamma,\phi|x)h(\gamma,\phi)}{p(\gamma,\phi)} p(\gamma,\phi) d\gamma d\phi}{\int \frac{L(\gamma,\phi|x)h(\gamma,\phi)}{p(\gamma,\phi)} p(\gamma,\phi) d\gamma d\phi} \end{aligned}$$

pode ser aproximado pela estimativa de Monte Carlo ponderado

$$\hat{H}(\gamma_* \mid x) = \frac{1}{n} \sum_{i=1}^{n} w_i \, I_{(-\infty,\gamma_*)}(\gamma_i), \tag{4.13}$$

cujos pesos são definidos por

$$w_i = \frac{L(\gamma_i, \phi_i \mid x) \, h(\gamma_i, \phi_i) / p(\gamma_i, \phi_i)}{\frac{1}{n} \sum_{j=1}^{n} L(\gamma_j, \phi_j \mid x) \, h(\gamma_j, \phi_j) / p(\gamma_j, \phi_j)}.$$

Denotando a amostra ordenada segundo γ_i por $\left(\gamma_{(i)}, \phi_{(i)}\right)$, $1 \le i \le n$, onde $\phi_{(i)}$ é o valor simulado concomitante de $\gamma_{(i)}$ (não é, pois, o i-ésimo menor valor de ϕ, em geral, no caso de ϕ ser também unidimensional), e por $w_{(i)}$ o peso respeitante ao par $\left(\gamma_{(i)}, \phi_{(i)}\right)$, a função de distribuição empírica ponderada de γ é então definida por

$$\hat{H}(\gamma_* \mid x) = \begin{cases} 0, & \gamma_* < \gamma_{(1)} \\ \sum_{j=1}^{i} w_{(j)}/n, & \gamma_{(i)} \le \gamma_* < \gamma_{(i+1)} \\ 1, & \gamma_* \ge \gamma_{(n)} \end{cases}$$

coincidindo naturalmente com a função de distribuição empírica usual quando $p(\gamma, \phi) = h(\gamma, \phi \mid x)$ (pois $w_{(i)} = 1$, $\forall i$).

Sendo γ_α o quantil de probabilidade α da distribuição marginal *a posteriori* de γ, i.e., $\gamma_\alpha = \inf\{\gamma : H(\gamma \mid x) \ge \alpha\}$, a sua estimativa Monte Carlo neste

contexto é

$$\hat{\gamma}_\alpha = \begin{cases} \gamma_{(1)}, & \alpha = 0 \\ \gamma_{(i)}, & \frac{1}{n}\sum_{j=1}^{i-1} w_{(j)} < \alpha \leq \frac{1}{n}\sum_{j=1}^{i} w_{(j)}. \end{cases}$$

Consequentemente, o intervalo central de credibilidade $1-\alpha$, $R_c(1-\alpha)$, é estimado (e de forma consistente) por

$$\hat{R}_c(1-\alpha) = \left(\hat{\gamma}_{\frac{\alpha}{2}}, \hat{\gamma}_{1-\frac{\alpha}{2}}\right). \tag{4.14}$$

Analogamente, denotando

$$\hat{R}_i(1-\alpha) = \left(\hat{\gamma}_{\frac{i}{n}}, \hat{\gamma}_{\frac{i+[n(1-\alpha)]}{n}}\right),\ i = 1, \ldots, n - [n(1-\alpha)]$$

uma sequência de intervalos de credibilidade $(1-\alpha)$ para γ, o intervalo HPD $R_0(1-\alpha)$ poderá ser estimado por $\hat{R}_0(1-\alpha) = \hat{R}_{i_0}(1-\alpha)$, onde $\hat{R}_{i_0}(1-\alpha)$ é o intervalo da referida sequência com a menor amplitude.

Havendo interesse em avaliar intervalos de credibilidade para funções reais $\psi(\gamma, \phi)$, basta calcular e ordenar os valores $\psi_i = \psi(\gamma_i, \phi_i)$ e estimar os seus quantis segundo o esquema mencionado, o que exige vantajosamente apenas o rearranjo dos mesmos pesos $w_{(i)}$.

4.2.2 Fatores de Bayes

Os intervalos de credibilidade são muitas vezes usados como meios de construir testes bayesianos de significância, como se referiu no Capítulo 1. Contudo, a adoção do método de Jeffreys implica que os testes de hipóteses (bem como os procedimentos usuais de comparação de modelos) se apoiem em fatores de Bayes, cuja determinação é geralmente complicada por serem razões de verosimilhanças marginais.

Concretamente, o fator de Bayes a favor de H_0 contra o modelo H_1 é a razão das verosimilhanças marginais $B(x) = p(x \mid H_0)/p(x \mid H_1)$ em que

$$p(x \mid H_k) = \int_{\Theta_k} f(x \mid \theta_k, H_k) h(\theta_k \mid H_k) d\theta_k,\ k = 0, 1.$$

Como tal, não é mais do que uma razão das constantes normalizadoras de $h(\theta_k \mid x, H_k)$, $k = 0, 1$, pelo que os métodos de avaliação de $B(x)$ a serem descritos são aplicáveis ao cálculo de qualquer quociente de constantes normalizadoras.

Começando por considerar o caso em que as densidades a integrar possuem a mesma dimensionalidade, seja $\theta_0 = \theta_1 \equiv \theta$ e tome-se para simplificar a notação,

$$h_k(\theta) \equiv h(\theta \mid x, H_k) = \bar{h}_k(\theta)/c_k,$$

com suporte Θ_k e $c_k = p(x \mid H_k)$, $k = 0, 1$. Pretende-se assim avaliar a razão

$$B(x) = \frac{c_0}{c_1} = \frac{\int_{\Theta_0} \bar{h}_0(\theta) d\theta}{\int_{\Theta_1} \bar{h}_1(\theta) d\theta}$$

quando os c_k são de difícil (ou impossível) avaliação analítica.

Denotando $p_k(\theta)$ a densidade de importância para $\bar{h}_k$, suposta completamente conhecida, geradora da amostra aleatória $\left(\theta_1^{(k)}, \ldots, \theta_{n_k}^{(k)}\right)$, $k = 0, 1$, a aplicação direta do método de Monte Carlo com amostragem de importância a cada c_k conduz à seguinte aproximação

$$\hat{B}_1(x) = \frac{\hat{c}_0}{\hat{c}_1}, \quad \hat{c}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{\bar{h}_k\left(\theta_i^{(k)}\right)}{p_k\left(\theta_i^{(k)}\right)}, \; k = 0, 1, \tag{4.15}$$

que é consistente para $B(x)$ (pelas leis dos Grandes Números).

Observe-se que o recurso ao método de Monte Carlo simples para a avaliação de c_k conduz à aproximação $\frac{1}{n_k} \sum_{i=1}^{n_k} L\left(\theta_i^{(k)} \mid x, H_k\right)$, onde os valores simulados representam agora uma amostra aleatória de dimensão n_k da distribuição *a priori* de θ sob H_k. Como esta pode ter um desempenho bastante diferenciado da verosimilhança, com a gama onde esta é apreciável pouco plausível *a priori*, os seus reflexos na amostra simulada implicam que esta aproximação de c_k pode ser bastante pobre. Veja-se o Cap. 5 para outros métodos de avaliação de c_k.

No caso particular de $\Theta_0 \subset \Theta_1$ é possível aplicar um procedimento de Monte Carlo para a avaliação de $B(x)$, recorrendo apenas a uma amostra gerada de $h_1(\theta)$ (possível de ser obtida com o desconhecimento de c_1 como se verá no Cap. 6). Com efeito,

$$B(x) = \frac{c_0}{c_1} = \int_{\Theta_1} \frac{\bar{h}_0(\theta)}{c_1} d\theta = E_{h_1}\left[\frac{\bar{h}_0(\theta)}{\bar{h}_1(\theta)}\right]$$

pelo que, sendo agora $\left(\theta_i^{(1)}, i = 1, \ldots, n_1\right)$ uma amostra aleatória de h_1,

$$\hat{B}_2(x) = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\bar{h}_0\left(\theta_i^{(1)}\right)}{\bar{h}_1\left(\theta_i^{(1)}\right)} \tag{4.16}$$

é uma estimativa de Monte Carlo (centrada e consistente) para $B(x)$. É de destacar que esta aproximação é muito eficiente quando as caudas de $h_1(\theta)$ são mais pesadas do que as de $h_0(\theta)$ e pouco eficiente quando há fraca sobreposição entre $h_0(\theta)$ e $h_1(\theta)$ (i.e., $E_{h_1}[h_0(\theta)]$ muito pequeno).

Em muitas situações, o fator de Bayes envolve funções densidade com dimensionalidade diferente - uma situação concreta respeita ao problema de testar hipóteses precisas na presença de parâmetros perturbadores através do método de Jeffreys. Para o tratamento deste problema mais geral vejam-se os livros Chen *et al.* (2000) e Paulino *et al.* (2003).

4.2.3 Densidades *a posteriori* marginais

Como se explicitou na Subsecção 4.1.3, a avaliação de densidades *a posteriori* marginais pela abordagem condicional de Gelfand e Smith (1990) exige o conhecimento completo (e não apenas o núcleo) das correspondentes densidades *a posteriori* condicionais, o que infelizmente não é a moeda corrente, e a disponibilidade de uma amostra da distribuição marginal *a posteriori* de $\theta^{(-m)}$. Se a obtenção desta é problemática e é possível arranjar para tal fim um sucedâneo $p(\cdot)$ de $h(\theta^{(-m)} \mid x)$ pode-se recorrer à estimativa Monte Carlo via amostragem de importância da densidade marginal de $\theta^{(m)}$

$$\hat{h}\left(\theta_*^{(m)} \mid x\right) = \frac{\sum_{i=1}^n w_i h\left(\theta_*^{(m)} \mid \theta_{(i)}^{(-m)}, x\right)}{\sum_{i=1}^n w_i} \tag{4.17}$$

onde $w_i = h\left(\theta_{(i)}^{(-m)} \mid x\right) / p\left(\theta_{(i)}^{(-m)}\right)$, $i = 1, \ldots, n$.

Observe-se que o conjunto de valores $\theta_{(i)}^{(m)}$, $i = 1, \ldots, n$, gerado pelo método de composição de $p\left(\theta^{(-m)}\right)$ e de $h\left(\theta^{(m)} \mid \theta^{(-m)}, x\right)$, não constitui uma amostra aleatória de $h\left(\theta^{(m)} \mid x\right)$ pelo facto de não se ter usado a distribuição *a posteriori* marginal de $\theta^{(-m)}$. Esses valores são visualizados como uma amostra ponderada de uma aproximação de $h\left(\theta^{(m)} \mid x\right)$, com pesos definidos por $p_i = w_i / \sum_{i=1}^n w_i$, $i = 1, \ldots, n$. Para obter dela uma amostra aleatória de dimensão L, diga-se, simulam-se no espírito do método *bootstrap* valores $\theta_{(j)*}^{(m)}$, $j = 1, \ldots, L$, da distribuição discreta $\left(\theta_{(i)}^{(m)}, p_i\right)$, $i = 1, \ldots, n$. Daí o nome de amostra *bootstrap* ponderado usado por Smith e Gelfand (1992).

Esta via de avaliar densidades *a posteriori* marginais, à semelhança da de Gelfand e Smith (1990), fica posta em causa quando a constante normaliza-

dora da densidade *a posteriori* condicional $h\left(\theta^{(m)} \mid \theta^{(-m)}, x\right)$ é desconhecida. Neste caso, Chen (1994) propõe um estimador ponderado de $h\left(\theta_*^{(m)} \mid x\right)$, com base na seguinte identidade

$$\begin{aligned} h\left(\theta_*^{(m)} \mid x\right) &= \int_{\Theta_{-m}\left(\theta_*^{(m)}\right)} h\left(\theta_*^{(m)}, \theta^{(-m)} \mid x\right) d\theta^{(-m)} \\ &= \int_{\Theta} w\left(\theta^{(m)} \mid \theta^{(-m)}\right) \frac{h\left(\theta_*^{(m)}, \theta^{(-m)} \mid x\right)}{h\left(\theta^{(m)}, \theta^{(-m)} \mid x\right)} h(\theta \mid x) d\theta, \end{aligned} \tag{4.18}$$

onde $\Theta_{-m}\left(\theta^{(m)}\right) = \left\{\theta^{(-m)} : \left(\theta^{(m)}, \theta^{(-m)}\right) \in \Theta\right\}$ é o subespaço de Θ fixado $\theta^{(m)}$ e $w\left(\theta^{(m)} \mid \theta^{(-m)}\right)$ é uma densidade condicional, completamente conhecida, com suporte igual ou contido no suporte de $h\left(\theta^{(m)} \mid \theta^{(-m)}, x\right)$, dado por $\Theta_m\left(\theta^{(-m)}\right) = \left\{\theta^{(m)} : \left(\theta^{(m)}, \theta^{(-m)}\right) \in \Theta\right\}$. Os critérios usados na seleção da função de importância são relevantes para a escolha da função w.

A forma da identidade (4.18) mostra que a constante normalizadora da distribuição *a posteriori* conjunta (e, *a fortiori*, condicional) cancela-se e que a densidade marginal em questão pode ser estimada sem viés com base numa amostra $\theta_{(i)} = \left(\theta_{(i)}^{(m)}, \theta_{(i)}^{(-m)}\right)$, $1 \leq i \leq n$, de $h(\theta \mid x)$, por

$$\hat{h}\left(\theta_*^{(m)} \mid x\right) = \frac{1}{n} \sum_{i=1}^{n} w\left(\theta_{(i)}^{(m)} \mid \theta_{(i)}^{(-m)}\right) \frac{h\left(\theta_*^{(m)}, \theta_{(i)}^{(-m)} \mid x\right)}{h\left(\theta_{(i)}^{(m)}, \theta_{(i)}^{(-m)} \mid x\right)}. \tag{4.19}$$

Capítulo 5

Avaliação de Modelos

Este capítulo debruça-se sobre procedimentos de avaliação de modelos tendo como finalidade operar uma seleção e comparação conducente à escolha do(s) melhor(es) modelo(s). Começa-se por descrever meios de diagnóstico ou de adequabilidade de modelos tais como valores-P bayesianos associados a diversas medidas de discrepância, resíduos, ordenadas preditivas condicionais e correspondentes fatores pseudo-Bayes. Em seguida, consideram-se algumas medidas de desempenho preditivo conhecidas pelos acrónimos AIC, SIC (ou BIC), DIC e WAIC e a forma de usar os fatores de Bayes para propósitos seletivos e comparativos. Por fim, dada a representação em modelos complexos da distribuição *a posteriori* por amostras dela simuladas, reportam-se meios de estimação de quantidades relevantes para a avaliação de modelos nesse contexto.

5.1 Crítica e adequabilidade de modelos

A verificação ou exame crítico de um modelo é a fase da análise estatística que visa avaliar a adequação do ajuste do modelo aos dados e ao que se sabe do problema em mão. Nesta fase pretende-se quantificar discrepâncias com os dados, avaliar se estas são ou não devidas ao acaso, averiguar o grau de sensibilidade das inferências para com os vários constituintes do modelo e descortinar vias de obter a partir dele modelos suscetíveis de serem mais

promissores.

Valores-P bayesianos

Box (1980, 1983) propôs para a análise crítica do modelo o confronto entre os dados x e a sua distribuição marginal (ou preditiva *a priori*) $p(x) = E_h[f(x|\theta)]$ através do valor-P marginal $P[p(X) \geq p(x)]$. Este procedimento é inaplicável em distribuições *a priori* impróprias, o que veda em particular o recurso à obtenção de réplicas dos dados observados a partir da simulação de $h(\theta)$. Mesmo quando $h(\theta)$ é própria, o método de Monte Carlo tende a ser instável em geral (Gelfand, 1996), e portanto pouco confiável, na avaliação de $p(x)$.

Uma das estratégias mais usadas de exame crítico do modelo assenta na distribuição preditiva *a posteriori* do modelo,

$$p(y|x) = \int_{\Theta} f(y|\theta)\, h(\theta|x)\, d\theta. \tag{5.1}$$

na base de que os dados y dela simuláveis devem refletir expectavelmente (ou não) os dados observados x, em caso de um bom (mau) ajustamento do modelo. Quaisquer desvios sistemáticos entre os dois conjuntos de dados, y e x, apontam para um virtual falhanço do modelo em subsequentes aplicações.

A medição da discrepância entre dados observados e observáveis de acordo com o modelo é traduzível por variáveis $V(x, \theta)$ cuja expressão depende naturalmente dos aspetos que se pretendem avaliar. Exemplos destas variáveis incluem desvios médios quadráticos padronizados definidos por

$$\sum_i \frac{[x_i - E(X_i|\theta)]^2}{Var(X_i|\theta)}$$

e a própria log-verosimilhança $\ln f(x|\theta)$, bem como os casos especiais traduzidos por estatísticas $V(x)$ e por funções paramétricas $g(\theta)$.

Como a distribuição preditiva *a posteriori* é tipicamente calculada por simulação da distribuição conjunta de (y, θ) condicional em x, considere-se então $\{(y_k, \theta_k),\ k = 1, \ldots, m)\}$ o conjunto de valores gerados dessa distribuição conjunta. O confronto entre os dados reais x e os dados preditos pelo modelo através da variável V pode evidenciar-se graficamente através, nomeadamente, do diagrama de dispersão dos valores $\{V(y_k, \theta_k), V(x, \theta_k),\ k = 1, \ldots, m\}$ ou do

histograma dos valores $\{V(y_k,\theta_k) - V(x,\theta_k),\, k = 1,\ldots,m\}$. Para um modelo bem ajustado aos dados, a nuvem de pontos do diagrama deve apresentar uma configuração simétrica em relação à bissetriz do 1º quadrante e o histograma deve incluir o valor 0.

Uma das medidas de resumo da discrepância entre $V(x,\theta)$ e a distribuição de $V(Y,\theta)$ é o valor-P preditivo *a posteriori*, por vezes apelidado de **valor-P bayesiano**, definido por

$$P_B = P\left[V(Y,\theta) \geq V(x,\theta)|x\right] \tag{5.2}$$

$$= E_{(Y,\theta)}\left\{I_{[V(x,\theta),+\infty)}[V(Y,\theta)]|x\right\}. \tag{5.3}$$

Note-se que um valor de P_B muito pequeno ou muito grande (diga-se inferior a 1% ou superior a 99%) traduz uma localização de $V(x,\theta)$ num ou noutro extremo da distribuição *a posteriori* condicional de $V(Y,\theta)$ dado x(sendo em regra relativamente implausível sob o modelo). Deste modo, ambos os casos evidenciam um mau ajuste do modelo aos dados no que concerne às características que a referida variável reflete. Também por isso o valor-P bayesiano pode, alternativamente, ser tomado como $P_B = P\left[V(Y,\theta) \leq V(x,\theta)|x\right]$, o que poderá ser visto como resultante da expressão anterior considerando o simétrico da respetiva variável. Tipicamente o valor-P bayesiano, P_B, é calculado como a proporção dos valores simulados da distribuição *a posteriori* de (Y,θ) dado x que verificam a condição $V(y_k,\theta_k) \geq V(x,\theta_k)$. Tais valores simulados obtêm-se por geração sucessiva de $h(\theta|x)$ e da distribuição amostral de cada observação vetorial futura condicionada (eventualmente) no vetor x de dados observados.

O exame preditivo *a posteriori* pode em alternativa ser aplicado separadamente por observação, fazendo uso das distribuições preditivas marginais $p(y_i|x)$. Os correspondentes valores-P marginais relativos à estatística $V(X_i)$ são definíveis por $\forall i,\ P_{B_i} = P\left[V(Y_i) \geq V(x_i)|x\right]$, convertendo-se em $P_{B_i} = P(Y_i \geq x_i|x)$ se V for a função identidade. Esta via é indicada para detetar observações discordantes e verificar se há uma correspondência global entre o modelo e os dados observados. Valores-P concentrados nos extremos ou no meio da sua gama de valores indicam, respetivamente, sobredispersão ou subdispersão relativa dos dados.

Valores extremos de uma medida de significância estatística como P_B não suscitam necessariamente um abandono categórico do modelo se a carate-

rística traduzida pela correspondente variável V não tiver grande relevância prática. Mesmo que a tenha, os aspetos sujeitos a crítica decorrentes desta inspecão do modelo devem servir como possíveis pistas de modificação no processo de procura de um modelo mais adequado à realidade dos dados concretos e do respetivo contexto.

Resíduos e outras medidas de diagnóstico/adequabilidade

Outras medidas podem ser construídas confrontando caraterísticas de distribuições preditivas do modelo baseadas em dados observados com outros dados igualmente observados. Isto é possível no contexto de validação cruzada em que haja uma amostra de treino $x = (x_1, \ldots, x_n)$ usada para gerar a partir do modelo a distribuição *a posteriori* $h(\theta|x)$, e uma amostra $y = (y_1, \ldots, y_l)$, independente de x, com o propósito de averiguar da validade do modelo em estudo à custa da correspondente distribuição preditiva *a posteriori* $p(y|x)$ ou de algumas das suas caraterísticas. Por exemplo, o valor médio e a variância preditivas de cada componente Y_j do vetor Y, de que y é uma concretização, são úteis para definir os *resíduos bayesianos preditivos* padronizados,

$$d_j = \frac{y_j - E(Y_j|x)}{\sqrt{var(Y_j|x)}}, \quad j = 1, \ldots, l, \tag{5.4}$$

os quais podem ser usados, à semelhança do que é feito na inferência clássica, como instrumento de averiguação informal da validade do modelo.

Esta via presume a existência de uma amostra de observações independentes, o que não acontece com frequência na prática. Claro que se a amostra inicial for de dimensão elevada, há sempre a possibilidade de a particionar em duas partes de modo que uma delas sirva de *amostra de treino* para construir a distribuição *a posteriori* e a outra de *amostra de validação* para obter a distribuição preditiva condicional à primeira.

Não sendo viável particionar a amostra global x para aplicar esta forma de validação cruzada, pode optar-se por um procedimento de tipo *jackknife* (*leave-one-out*) que consiste em repetir n vezes o esquema de validação cruzada deixando de fora das subamostras de treino um ponto observacional de cada vez, o qual desempenhará o papel de subamostra de validação. Daí a designação de procedimento de **validação cruzada com um de fora**.

Assim, se se designar por $x_{(-i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, o vetor consti-

tuído por todas as observações à exceção de x_i, pode obter-se as *distribuições preditivas condicionais* $p(y_i|x_{(-i)})$,

$$p(y_i|x_{(-i)}) = \int f(y_i|\theta, x_{(-i)})\ h(\theta|x_{(-i)})\ d\theta, \tag{5.5}$$

e os consequentes *resíduos bayesianos de eliminação* padronizados

$$d_i' = \frac{x_i - E(Y_i|x_{(-i)})}{\sqrt{var(Y_i|x_{(-i)})}}, \quad i = 1, \ldots, n, \tag{5.6}$$

onde os valores médios e variâncias são calculados, analiticamente ou por simulação, das correspondentes distribuições preditivas condicionais.

Pode proceder-se novamente a uma validação informal do modelo à custa destes resíduos. Por outro lado, os valores calculados em x_i de $p(y_i|x_{(-i)})$, comummente designados por **ordenadas preditivas condicionais** (CPO), podem ainda ser usados num diagnóstico informal. Com efeito, estes valores são um indicador da verosimilhança de cada observação dadas todas as outras observações e, portanto, valores baixos de CPO devem corresponder a observações mal ajustadas. Neste sentido, um modelo será tanto mais adequado quanto maior for a soma dos logaritmos das ordenadas preditivas condicionais

$$\sum_{i=1}^{n} \ln CPO_i = \ln \prod_{i=1}^{n} p(x_i|x_{(-i)}). \tag{5.7}$$

Para outros instrumentos de diagnóstico veja-se designadamente Gelfand (1996) e Gelman e Meng (1996).

Exemplo 5.1 Um estudo do desempenho de modelos de automóvel medido pelo consumo de combustível está descrito em Henderson e Vellman (1981), do qual se coligiu um subconjunto de dados. Estes dados reportam-se aos valores de eficiência, E_f, medida em milhas percorridas por galão de gasolina, peso em libras (X_1), potência em cavalos-vapor (X_4^*) e número de marchas da caixa de velocidades nos níveis 3, 4 e 5 representado conjuntamente pelas variáveis indicadoras das categorias 4 (X_2) e 5 (X_3). Na sequência de trabalho preliminar, consideraram-se modelos de regressão Normal na variável resposta transformada $Y = 100/E_f$, expressa em galões consumidos por 100 milhas.

Um dos modelos considerados envolve as variáveis explicativas X_1, (X_2, X_3) e $X_4 = X_4^*/X_1$ (potência por unidade de peso) através da função de regressão

múltipla

$$M_1 : \mu \equiv E(Y) = \beta_0 + \sum_{j=1}^{4} \beta_j X_j + \beta_5 X_2 X_4 + \beta_6 X_3 X_4,$$

que corresponde a 3 funções de regressão distintas em X_1 e X_4, uma para cada número de velocidades, diferindo entre si na ordenada na origem e no declive com X_4.

O modelo de regressão $Y_i, i = 1, \ldots, n = 29 \overset{iid}{\sim} N(\mu, \sigma^2)$ foi complementado com distribuições *a priori* não informativas do tipo usual, especificamente $\beta_j \sim N(0, 10^4)$ e $1/\sigma^2 \sim Ga(10^{-3}, 10^{-3})$. A análise bayesiana deste modelo linear sob uma distribuição *a priori* conjugada natural ou a distribuição não informativa usual para (μ, σ^2) é em muitos aspetos obtenível analiticamente em termos exatos (*vide, e.g.*, Paulino *et al.*, 2003, Sec. 4.3).

Para ilustração de quantidades descritas neste capítulo sob o modelo bayesiano acima e algumas reduções dele, com μ parametrizado em termos dos coeficientes de regressão, seguiu-se a via da simulação por métodos de Monte Carlo. Os modelos reduzidos que vão competir com o modelo indicado obtêm-se deste simplificando μ por retirada dos termos de interação (M_2) e também dos efeitos principais das variáveis binárias representando o número de marchas (M_3). Denota-se por θ o vetor de parâmetros, constituído pelos coeficientes de regressão e variância, de cada modelo.

A Figura 5.1 apresenta os diagramas de dispersão para os modelos M_1 e M_2 da medida de discrepância traduzida pelos desvios médios quadráticos padronizados, $V(\cdot, \theta_k)$, calculados de valores simulados θ_k da distribuição *a posteriori* de θ (dados os valores observados de Y e $\{X_j\}$) para os dados reais *versus* dados preditos.

A nuvem de pontos deste tipo de diagramas está em ambos os casos mais concentrada na região acima da bissetriz do 1º quadrante, indicando que os referidos desvios preditos pelos dois modelos tendem a ser maiores do que os correspondentes desvios para os dados observados. A configuração assimétrica desta nuvem em relação à bissetriz parece um pouco menos pronunciada para M_2 do que para M_1. Os valores-P bayesianos associados com a discrepância entre a função V avaliada nos dados observados e a distribuição *a posteriori* de $V(Y^*, \theta)$ para dados predizíveis Y^*, para os 3 modelos registados na Tabela 5.1, apontam no sentido em que os modelos reduzidos se comportam melhor do que o modelo encaixante em termos do grau de ajuste aos dados. A

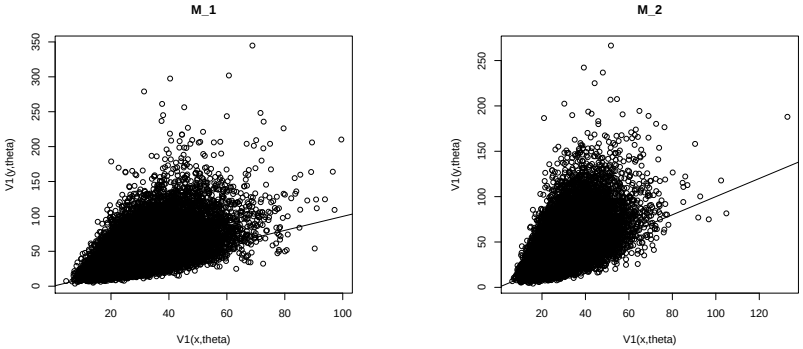

Figura 5.1: Diagramas de dispersão para os modelos M_1 e M_2

soma dos logaritmos das ordenadas preditivas condicionais aponta no mesmo sentido com uma ligeiríssima vantagem para o modelo M_2. Estas quantidades foram estimadas segundo a sugestão de Gelfand (1996) pela média harmónica das densidades amostrais avaliadas nos valores de uma amostra simulada da distribuição *a posteriori* de θ (*vide* Subsecção 5.3.1).

Tabela 5.1: Medidas de diagnóstico dos modelos em comparação

Modelo	P_B	$\sum \ln CPO$
M_1	0.844	-27.577
M_2	0.778	-23.665
M_3	0.687	-23.817

■

5.2 Seleção e comparação de modelos

A avaliação dos modelos estatísticos pode fazer-se através de vários critérios cuja relevância relativa depende dos propósitos que se tem em mente quando se analisam os dados entretanto observados recorrendo a um ou mais modelos escolhidos para o efeito, como aliás se deixou transparecer na descrição prévia da fase de verificação crítica dos modelos. Ora parece pacífico que uma ideia

central na avaliação de um modelo seja a precisão de predições obteníveis do modelo.

A medida ideal do ajuste do modelo deve refletir o seu desempenho preditivo num quadro de validação externa, i.e., em relação a novos dados advenientes do verdadeiro modelo que é naturalmente desconhecido. Denotando estes dados hipotéticos por $y = (y_i)$ e os dados já observados por $x = (x_i)$, uma medida de acurácia preditiva para cada y_i poderia ser o logaritmo da sua distribuição preditiva *a posteriori* $\ln p(y_i|x) = \ln E_{\theta|x}[f(y_i|\theta)]$.

Dado o desconhecimento do modelo que gera os dados e o caráter fictício de $\{y_i\}$, consegue-se uma aproximação daquela medida tomando no lugar destes os dados reais, obtendo-se então a medida de acurácia preditiva intra-amostra

$$\tilde{A}(x) = \sum_{i=1}^{n} \ln p(x_i|x) = \ln \prod_{i=1}^{n} p(x_i|x) \tag{5.8}$$

$$\simeq \sum_{i=1}^{n} \ln \left[\frac{1}{m} \sum_{k=1}^{m} f(x_i|\theta_k) \right], \tag{5.9}$$

onde θ_k, $k = 1, \ldots, m$ são valores simulados da distribuição *a posteriori* $h(\theta|x)$.

Este tipo de aproximação da medida de acurácia preditiva extra-amostra é fácil de obter mas tende a sobrestimar tal medida em geral porque envolve os dados observados duplamente, na atualização distribucional do parâmetro e no uso da distribuição amostral para cada elemento dos dados. Este aspeto sugere o recurso a correções a $\tilde{A}(x)$, o que complementado com a aplicação de transformações dá origem a distintas medidas de desempenho preditivo, tradicionalmente alcunhadas de critérios de informação.

5.2.1 Medidas de desempenho preditivo

Descrevem-se a seguir algumas destas medidas que têm merecido o maior interesse no plano prático ou teórico. Pelo modo como os critérios de informação são todos definidos, especialmente negativando as correspondentes medidas de acurácia, o desempenho do modelo será tanto melhor quanto mais baixos forem os valores desses critérios, como se evidenciará em seguida.

Critério de informação de Akaike (AIC)

A medida de acurácia preditiva usada neste critério devido a Akaike (1973) assenta na aproximação de $\tilde{A}(x)$ pela log-verosimilhança maximizada, $\ln f(x|\hat{\theta})$, segundo uma estratégia de enfiamento (*plug-in*) bem típica da Estatística Frequencista, descontada do número p de parâmetros representados em θ, i.e. por

$$\tilde{A}_{AIC} = \ln f(x|\hat{\theta}) - p, \tag{5.10}$$

onde $\hat{\theta}$ é a estimativa MV de θ. A medida de informação AIC obtém-se de $\tilde{A}_{AIC}$ por uma transformação linear idêntica à da estatística da razão de verosimilhanças de Wilks, ou seja

$$AIC = -2 \ln f(x|\hat{\theta}) + 2p, \tag{5.11}$$

mas qualquer constante negativa a multiplicar a medida de acurácia preditiva terá o mesmo efeito. Em termos práticos esta medida compõe-se assim de duas componentes, uma associada à qualidade de ajuste do modelo e a outra à complexidade deste quantificada por duas vezes a dimensão paramétrica. Este critério pela sua expressão não depende da dimensão amostral mas a sua derivação pressupõe de facto um contexto de grandes amostras. Por outro lado, toda a pouca ou muita informação apriorística sobre os parâmetros do modelo amostral é simplesmente ignorada pelo AIC em contraponto com critérios bayesianos que se descrevem em seguida. Para informações mais detalhadas sobre este critério originariamente alheio à metodologia bayesiana, nomeadamente sobre variantes suas,*vide* Burnham and Anderson, 2002, caps. 2 e 6.

Critério de informação Bayes (SIC/BIC)

Schwarz (1978) propõe um critério (SIC) que permite selecionar modelos através de uma aproximação para grandes amostras da respetiva distribuição marginal dos dados $p(x) = E_{h(\theta)}[f(x|\theta)]$. Tal critério de informação tem sido mais frequentemente rotulado por BIC (de *Bayes Information Criterion*), possivelmente por se basear na ponderação da distribuição amostral dos dados pela densidade *a priori*, refletindo uma génese bem distinta daquela ligada a medidas de acurácia preditiva *a posteriori*.

Como resultado da seguinte aproximação, para grandes valores do tamanho amostral n, $\ln p(x) \simeq \ln f(x|\hat{\theta}) - (p/2) \ln n$, onde $\hat{\theta}$ é a estimativa MV de θ, o

critério BIC é definido por

$$BIC = -2 \ln f(x|\hat{\theta}) + p \ln n. \tag{5.12}$$

Este critério considera preferíveis os modelos com maiores valores da referida aproximação de $p(x)$, ou equivalentemente, menores valores de BIC. A parte relativa à dimensão do modelo no BIC é bem maior do que a do AIC em amostras moderadas ou grandes, deste modo penalizando mais os modelos mais complicados.

De modo a obviar ao cálculo de máximos por simulação, Carlin e Louis (2000) sugerem uma modificação a este critério usando a média *a posteriori* da log-verosimilhança em vez do seu máximo. Esta versão modificada do BIC fica então definida por

$$BIC^{CL} = -2E_{\theta|x} \left[\ln f(x|\theta) \right] + p \ln n. \tag{5.13}$$

Critério de informação pela desviância (DIC)

Tendo em consideração que a distribuição *a priori* e o tipo de estrutura do modelo tendem a afetar o grau de sobreajustamento, o critério DIC proposto por Spiegelhalter *et al.* (2002) corresponde a modificar a medida de acurácia preditiva $\tilde{A}_{AIC}$ em dois aspetos ligados às suas duas componentes. A estimativa MV $\hat{\theta}$ é substituída por uma estimativa bayesiana $\bar{\theta}$ (frequentemente a média *a posteriori* de θ) e o número de parâmetros por uma dimensão efetiva do modelo, p_D, definida como se argumenta em seguida.

Para a definição da complexidade do modelo, Spiegelhalter *et al.* (2002) partem de uma medida de informação relativa na distribuição amostral definida por $-2 \ln [f(x|\theta)/g(x)]$, onde $g(x)$ denota alguma função apenas dos dados com efeito meramente padronizador. Por exemplo, $g(x) = f(x|\tilde{\theta})$ em que $\tilde{\theta}$ é a estimativa de θ no modelo saturado, o que faz com que aquela medida de informação possa ser entendida como a função paramétrica associada com a estatística da razão das verosimilhanças. Vista como função de θ para os dados x, aquela medida é frequentemente chamada **desviância** (de *deviance*) bayesiana e denotada por $D(\theta)$.

Em consonância, o valor dessa medida de informação relativa na mesma distribuição quando θ é estimado por $\bar{\theta}$ é denotada como $D(\bar{\theta})$ e, deste modo,

a quantidade de informação na distribuição amostral que a estimação de θ por $\bar{\theta}$ não consegue tomar em conta é descrita pela diferença

$$D(\theta) - D(\bar{\theta}) = -2 \ln \frac{f(x|\theta)}{f(x|\bar{\theta})}, \tag{5.14}$$

a qual já não depende do termo padronizador $g(x)$. Como θ é desconhecido e, como tal, bayesianamente aleatório, Spiegelhalter *et al.* (2002) substituem $D(\theta)$ pelo seu valor esperado *a posteriori*, $\overline{D(\theta)} = E_{\theta|x}[D(\theta)]$, considerando o **número efetivo de parâmetros** como sendo definido por

$$p_D = \overline{D(\theta)} - D(\bar{\theta}), \tag{5.15}$$

ou seja, como a diferença entre a desviância média *a posteriori* e a desviância na média *a posteriori* quando $\bar{\theta} = E(\theta|x)$. Naturalmente que esta quantidade depende dos dados, do foco paramétrico e da informação *a priori*.

Outras estimativas bayesianas de θ e $D(\theta)$ poderão ser usadas, em alternativa, com resultados naturalmente distintos dos que se obtêm com o uso de médias. Uma das vantagens da definição acima está na trivial calculabilidade desde que o modelo possua uma verosimilhança em forma fechada, dado que

$$p_D \simeq \frac{1}{m} \sum_{k=1}^{m} D(\theta_k) - D(\frac{1}{m} \sum_{k=1}^{m} \theta_k), \tag{5.16}$$

onde os θ_k são valores simulados de $h(\theta|x)$. Outra vantagem reside na garantia de $p_D \geq 0$ para qualquer modelo com verosimilhança log-côncava (pela desigualdade de Jensen).

Em vários modelos de tipo padrão (sem uma estrutura hierárquica) com uma verosimilhança dominando a distribuição *a priori*, verifica-se que p_D aproxima-se do número real de parâmetros. A proposta de p_D como uma medida de dimensionalidade do modelo visa alargar a sua aplicabilidade a modelos complexos em que não é fácil determinar o número real de parâmetros como os modelos com variáveis latentes. Todavia, a ineficácia dessa medida decorrente da obtenção de valores negativos para p_D é um facto constatado em alguns desses modelos como os estruturados em misturas finitas ou hierarquias (veja-se, por exemplo, Celeux *et al.*, 2006).

Tomando p_D como uma medida de complexidade do modelo, a correspondente medida de acurácia preditiva dada então por

$$\tilde{A}_{DIC} = \ln f(x|\bar{\theta}) - p_D, \tag{5.17}$$

pode então traduzir-se no critério DIC (de *deviance information criterion*) proposto em Spiegelhalter *et al.* (2002) por

$$DIC = D(\bar{\theta}) + 2p_D = \overline{D(\theta)} + p_D = 2\overline{D(\theta)} - D(\bar{\theta}). \tag{5.18}$$

Na prática, costuma-se omitir da medida DIC o termo padronizante da desviância (o que equivale a tomar $g(x) = 1$) quando os modelos bayesianos a comparar assentam no mesmo modelo amostral, embora diferindo na estrutura paramétrica. De outro modo, é preciso ser cauteloso porque a medida DIC depende de tal função.

Critério de informação amplamente aplicável (WAIC)

A medida de acurácia preditiva intra-amostra associada ao critério WAIC (de *widely applicable information criterion*), devido a Watanabe (2010), não é aproximada, como nos critérios anteriores, segundo um esquema de enfiamento, sendo por isso mais bayesiana que a usada pelo DIC e expressa, quando corrigida, por

$$\tilde{A}_{WAIC} = \tilde{A}(x) - p_W = \sum_{i=1}^{n} \ln E_{\theta|x}\left[f(x_i|\theta)\right] - p_W. \tag{5.19}$$

Tal como $\tilde{A}(x)$, o termo corretivo do sobreajustamento também se baseia em cálculos para cada um dos pontos observacionais cujos resultados são depois adicionados. Uma das propostas para p_W assemelha-se à usada no DIC, sendo respetivamente definida e calculada por

$$p_{W_1} = -2\sum_{i=1}^{n}\{E_{\theta|x}\left[\ln f(x_i|\theta)\right] - \ln E_{\theta|x}\left[f(x_i|\theta)\right]\} \tag{5.20}$$

$$\simeq -2\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{k=1}^{m}\ln f(x_i|\theta_k) - \ln\left[\frac{1}{m}\sum_{k=1}^{m} f(x_i|\theta_k)\right]\right\}, \tag{5.21}$$

fórmulas estas que evidenciam que $p_{W_1} \geq 0$.

Outra proposta para p_W baseia-se na variância *a posteriori* de $\ln f(x_i|\theta)$ para todos os dados, sendo definida e calculada respetivamente por

$$p_{W_2} = \sum_{i=1}^{n} Var_{\theta|x}\left[\ln f(x_i|\theta)\right] \tag{5.22}$$

$$\simeq \sum_{i=1}^{n}\left\{\frac{1}{m-1}\sum_{k=1}^{m}\left[l_k(x_i) - \bar{l}(x_i)\right]^2\right\}, \tag{5.23}$$

em que $l_k(x_i) = \ln f(x_i|\theta_k)$ e $\bar{l}(x_i) = \frac{1}{m}\sum_{k=1}^{m} l_k(x_i)$.

Usando qualquer destas sugestões de termos corretivos de $\tilde{A}(x)$, vistos como "dimensão efetiva do modelo" dependente quer dos próprios dados quer da distribuição *a priori*, Gelman *et al.* (2014 a,b) propõem uma transformação do critério de informação de Watanabe na mesma escala da desviância para propósitos comparativos expressável por

$$WAIC = -2\sum_{i=1}^{n} \ln E_{\theta|x}\left[f(x_i|\theta)\right] + 2p_W. \tag{5.24}$$

5.2.2 Seleção por comportamento preditivo *a posteriori*

A análise bayesiana típica repousa num dado modelo e, como tal, as respetivas inferências revestem uma natureza condicional. Ora as incertezas inevitavelmente presentes em qualquer das componentes do modelo bayesiano são razões suficientemente fortes para se contemplar uma gama possível de modelos conjuntos para os dados e parâmetros, a qual deve ser sujeita a análises para uma triagem preliminar.

Após a inspeção e avaliação dos modelos, deve-se proceder à sua comparação visando a seleção dos que melhor se comportam segundo os critérios adotados. Esta fase da análise estatística não tem que culminar necessariamente na escolha de um único modelo porque tal pode levar a conclusões enganadoras. Ao invés, deve-se certamente descartar alguns modelos com desempenho inaceitável à luz dos critérios escolhidos e guardar os restantes para ulterior consideração[14].

A prática tradicional de fazer seleção de modelos estatísticos através de testes de hipóteses tem vindo a mudar desde há 40 anos, possivelmente pela consciência crescente das limitações desses procedimentos quando confrontados com métodos baseados na otimização de medidas de desempenho.

O modo de aplicação dos critérios de informação anteriormente descritos respeita o princípio da parcimónia que estipula que os modelos a escolher devem aliar uma representação adequada dos dados (uma boa qualidade do ajuste) a uma estrutura tão simples quanto possível (dimensão pequena). Este

[14]Isto pode inclusivamente respeitar à realização de inferências de interesse através de apropriada combinação dos modelos selecionados, conhecida como ponderação bayesiana de modelos (*Bayesian Model Averaging*).

princípio exprime a máxima bem remota (século XIV) da **rasoura de Occam**, *shave away all that is unnecessary*.

Tendo em vista a necessidade de reduzir a cardinalidade do conjunto de modelos a comparar, uma estratégia sensata de aplicação desse princípio deve pautar-se pelas seguintes ideias: descartar o modelo que prediz os dados ou menos bem que os modelos nele encaixados ou muito pior do que aquele que correntemente produz as melhores predições. Esta estratégia é naturalmente compatível com diferentes modos de medir a qualidade de predição, sejam eles instrumentos de diagnóstico, critérios de informação ou outros.

Uso de medidas de diagnóstico

Em face de vários modelos em competição, medidas de diagnóstico ou de adequabilidade como o valor-P bayesiano e os resíduos bayesianos padronizados (quer obtidos por validação cruzada, quer por *jackknife*) são úteis para avaliar comparativamente tais modelos em termos do seu desempenho. Por exemplo, usando a soma dos quadrados (ou dos valores absolutos) desses resíduos deve-se privilegiar os modelos que apresentam os menores valores.

Um outro método para avaliar a adequabilidade comparativa entre vários modelos consiste em calcular a soma dos logaritmos das ordenadas preditivas condicionais, tendo em vista escolher os modelos que apresentem os maiores valores. Quando se pretende comparar dois modelos H_1 e H_2 através das CPO pode usar-se a quantidade

$$\prod_{i=1}^{n} \frac{p(x_i|x_{(-i)}, H_1)}{p(x_i|x_{(-i)}, H_2)} \equiv \prod_{i=1}^{n} Q_i, \tag{5.25}$$

denominada *fator pseudo-Bayes* na base de o produto das CPO para cada modelo ser proposto como "substituto" da respetiva verosimilhança marginal - recorde-se a definição do fator Bayes. Se este fator for maior do que 1, prefere-se H_1 a H_2, em caso contrário prefere-se H_2. Além disso, como observações para as quais o termo Q_i é maior (menor) que 1 indicam preferência pelo modelo H_1 (H_2), um gráfico de $\ln(Q_i)$ versus i é útil para uma visão global das observações que se encontram melhor ajustadas para cada um dos modelos em competição.

Uso de critérios de informação

No cenário de comparação de modelos o que é relevante não são os valores das medidas de desempenho preditivo em termos absolutos mas antes os valores relativos dentro do conjunto de J (diga-se) modelos em competição. Daí a utilidade das diferenças entre os valores de cada uma dessas medidas para pares de modelos. Como as medidas consideradas na subsecção anterior, e que por agora se denotam genericamente por IC, são todas consonantes com a atribuição dos seus menores valores aos melhores modelos, torna-se útil determinar para cada modelo as diferenças $r_j(IC) = IC_j - IC_{min}$, $j \in J$.

Calcular estas diferenças, que operam uma nova escala, permite uma mais fácil comparação e ordenação dos modelos em consideração. Quanto menor for r_j tanto maior é o grau de sustentação empírica do modelo j, passando o melhor modelo da classe em questão (que aqui se vai considerar indexado por o, de ótimo) a ser indicado pelo valor 0 dessa diferença.

Como exemplificação de diferenças r_j, eis a sua expressão para a medida DIC:

$$\begin{aligned} r_j(DIC) &= DIC_j - DIC_o \\ DIC_j &= E_{\theta_j|x}[D_j(\theta_j)] + p_{D_j} = 2E_{\theta_j|x}[D_j(\theta_j)] - D_j(E_{\theta_j|x}[\theta_j]). \end{aligned}$$

Convém notar em relação às diferenças respeitantes ao critério BIC, em que se denotam as dimensões paramétricas momentaneamente por $\{p_j^*\}$, que

$$\begin{aligned} r_j(BIC) = BIC_j - BIC_o &= -2\ln \frac{f_j(x|\hat{\theta}_j)n^{-p_j^*/2}}{f_o(x|\hat{\theta}_0)n^{-p_o^*/2}} \\ &\simeq -2\ln \frac{p_j(x)}{p_o(x)} = -2\ln B_{jo}(x). \end{aligned} \tag{5.26}$$

Isto é, as diferenças $r_j(BIC)$ estão relacionadas com o fator Bayes entre os modelos em confronto, diga-se H_j e H_o, no sentido de serem aproximadas para grandes amostras por $-2\ln B_{jo}(x)$, segundo o argumento de Schwarz (1978).

Exemplo 5.2 Continuando com o exemplo anterior para efeitos de avaliação comparativa de 3 modelos de regressão múltipla em termos do seu comportamento preditivo, indicam-se em seguida os fatores pseudo-Bayes (PBF)

relativos à comparação dos modelos dois a dois:

$$PBF(M_1/M_2) = 0,809; PBF(M_1/M_3) = 0,941; PBF(M_2/M_3) = 1,164.$$

Estes valores mostram bem as constatações feitas anteriormente de que M_2 é o melhor enquanto M_1 é o pior dos 3 modelos, em termos do critério assente nas ordenadas preditivas condicionais.

Em termos dos critérios de informação exibidos na Tabela 5.2, esta mostra que o modelo M_2 é o melhor modelo em termos das medidas bayesianas DIC e WAIC, sendo batido por M_3 no BIC, sem grande espanto uma vez que se sabe que esta medida beneficia os modelos mais simples.

Tabela 5.2: Medidas DIC, BIC e WAIC para os modelos em comparação

Modelo	DIC (p_D)	BIC (p)	WAIC (p_{W_2})
M_1	48.69 (8.27)	67.36 (8)	46.77 (5.38)
M_2	47.32 (6.19)	61.33 (6)	46.70 (4.78)
M_3	47.58 (4.12)	56.93 (4)	47.40 (3.48)

Conjugando todos os resultados obtidos aqui e no exemplo anterior, os critérios aplicados indicam que o melhor dos três modelos é aquele que se situa no meio da escala de complexidade medida pelo número de parâmetros.

■

5.2.3 Seleção via Fator de Bayes

Os métodos de escolha de modelos mencionados na subsecção anterior procedem a uma triagem no sentido de reter para consideração ulterior modelos que se verifica apresentarem um bom desempenho à luz da informação recolhida, sem se preocuparem se algum dos modelos representa a realidade em estudo.

Em contraponto, existem métodos cuja formalização os enquadra numa perspetiva (bem controversa) de que a classe dos modelos em análise integra o modelo dito verdadeiro, na aceção de ser o responsável pelos dados obtidos (ou, pelo menos, um que constitua uma aproximação suficientemente razoável daquele). O método dos fatores de Bayes para efeitos de seleção no quadro de

uma classe discreta de modelos insere-se precisamente nessa discutível perspetiva. Este e outros aspetos criticáveis não têm impedido que este método, a ser descrito em seguida, seja utilizado em alguns contextos e com as devidas precauções.

Uma gama geral de modelos bayesianos (distintos modelos amostrais com distintas distribuições *a priori*) origina uma classe correspondente de modelos preditivos *a priori*, $\mathcal{M} = \{M_j,\ j \in J\}$, cujas distribuições são definíveis por

$$p(x \mid M_j) \equiv p_j(x) = \int f_j(x \mid \theta_j) h_j(\theta_j) d\theta_j, \quad j \in J\ . \tag{5.27}$$

Sendo J um conjunto discreto de rótulos indexadores dos membros da classe que se admite conter o desconhecido modelo verdadeiro, a distribuição preditiva global corresponde a

$$p(x) = \sum_{j \in J} P(M_j) p(x \mid M_j)\ , \tag{5.28}$$

onde $P(M_j)$ é a probabilidade *a priori* de M_j ser o modelo verdadeiro que é, pois, atualizada em face dos dados x para

$$P(M_j \mid x) = P(M_j) p(x \mid M_j)/p(x), \quad j \in J\ . \tag{5.29}$$

O fator de Bayes a favor de M_k e contra M_j é então a razão das chances

$$B_{kj}(x) = \frac{P(M_k \mid x)/P(M_j \mid x)}{P(M_k)/P(M_j)} = \frac{p(x \mid M_k)}{p(x \mid M_j)}\ , \tag{5.30}$$

sendo as chances *a posteriori*de M_k expressáveis por

$$\frac{P(M_k \mid x)}{1 - P(M_k \mid x)} = \frac{P(M_k)}{\sum_{j \neq k} P(M_j) B_{jk}(x)}, \tag{5.31}$$

em que $B_{jk}(x) = 1/B_{kj}(x)$ é o fator de Bayes a favor de M_j e contra M_k, especializando-se essa fórmula em $(\sum_{j \neq k} B_{jk}(x))^{-1}$ quando se usa no espaço dos modelos uma distribuição *a priori* uniforme. Exemplos simples de aplicação deste método podem encontrar-se em Paulino *et al.* (2003).

Na generalidade das situações, a determinação das densidades preditivas *a priori* e, em decorrência, dos fatores de Bayes é feita recorrendo a simulação - veja-se para o efeito o disposto no capítulo precedente e a secção seguinte. A comparação dos modelos por pares através do fator de Bayes requer o uso de regras práticas que especifiquem limiares delimitadores dos níveis qualitativos da força de evidência a favor de um modelo contra o seu oponente. Uma dessas regras é avançada por Kass e Raftery (1995).

5.3 Simulação em avaliação de modelos

Na sequência do Cap. 4, as secções 5.1 e 5.2 deram já indicações concretas sobre como recorrer a meios de simulação estocástica para determinação por métodos de Monte Carlo de medidas de diagnóstico, adequação e desempenho preditivo dos modelos estatisticos. Nesta secção complementa-se esse material com questões adicionais relevantes para a execução do cálculo bayesiano por simulação.

Para o efeito, começa-se por fixar o contexto de se dispor de uma amostra $\{\theta^{\star}_{(j)};\ j = 1,\ldots,m\}$ da distribuição *a posteriori* $h(\theta|x)$, obtida previamente por algum método de simulação. Em avaliação de modelos interessa nomeadamente resolver problemas que se mencionam nas subsecções seguintes.

5.3.1 Estimação de densidades preditivas *a posteriori*

Se y representar um novo conjunto de dados independente daquele já observado, a densidade preditiva $p(y|x)$ pode ser simplesmente estimada como se referiu na Subsecção 4.1.4 por

$$\hat{p}(y|x) = \frac{1}{m}\sum_{j=1}^{m} f(y|\theta^{\star}_{(j)}).$$

Seguindo o mesmo raciocínio, para a estimação da densidade preditiva condicional dada por

$$p(x_i|x_{(-i)}) = \int f(x_i|\theta, x_{(-i)})\ h(\theta|x_{(-i)})\ d\theta,$$

ter-se-ia, em princípio, de simular amostras $\{\theta^{\star}_{(j)};\ j = 1,\ldots,m\}$ para cada distribuição *a posteriori* $h(\theta|x_{(-i)})$. Ora isto não é de modo algum prático quando x tem dimensão elevada. O objetivo é contornar o problema e fazer a estimação das densidades pretendidas à custa meramente da amostra $\{\theta^{\star}_{(j)};\ j = 1,\ldots,m\}$ de $h(\theta|x)$.

Gelfand (1996) sugere a utilização da média harmónica de $\{f(x_i|x_{(-i)},\theta^{\star}_{(j)}),\ j = 1,\ldots,m\}$ para estimar $p(x_i|x_{(-i)})$. Com efeito, considerando que

$$p(x)h(\theta|x) = h(\theta)f(x|\theta) = h(\theta)f(x_i|x_{(-i)},\theta)f(x_{(-i)}|\theta),$$

tem-se

$$p(x_i|x_{(-i)}) = \frac{p(x)}{p(x_{(-i)})} = \frac{1}{\int \frac{1}{f(x_i|x_{(-i)},\theta)}\ h(\theta|x)\ d\theta},$$

e portanto, se $\{\theta^{\star}_{(j)}; j = 1, \ldots, m\}$ é uma amostra de $h(\theta|x)$, tem-se

$$\hat{p}(x_i|x_{(-i)}) = \frac{1}{\frac{1}{m}\sum_{j=1}^{m} \frac{1}{f(x_i|x_{(-i)}, \theta^{\star}_{(j)})}}. \tag{5.32}$$

Esta expressão simplifica-se no caso em que $(X_1, \ldots, X_n)$ são condicionalmente independentes, já que nessa situação, $f(x_i|x_{(-i)}, \theta) = f(x_i|\theta)$.

Note-se que o mesmo tipo de argumentos permite estimar momentos da distribuição preditiva *a posteriori* (e, em particular, de resíduos bayesianos preditivos padronizados) à custa de momentos amostrais avaliados em cada ponto da amostra simulada de $h(\theta|x)$, bem como outras medidas de diagnóstico.

5.3.2 Estimação da densidade preditiva *a priori*

A obtenção da densidade preditiva *a priori* $p(x)$, quando ela existe, é necessária para o cálculo de fatores de Bayes. Passam-se em revista algumas das propostas surgidas na literatura para a sua obtenção.

Se $\{\theta^{\star}_{(j)}; j = 1, \ldots, m\}$ for uma amostra da distribuição *a priori* $h(\theta)$, decorre da definição de $p(x)$ que esta pode ser aproximada por

$$\hat{p}(x) = \frac{1}{m}\sum_{j=1}^{m} f(x|\theta^{\star}_{(j)}).$$

Esta estimativa de $p(x)$ tem-se revelado, contudo, muito ineficiente em geral como se referiu na Subsecção 4.2.2.

Newton e Raftery (1994) sugerem usar a média harmónica de $\{f(x|\theta^{\star}_{(j)})\}$, ou seja

$$\hat{p}(x) = \Big[\frac{1}{m}\sum_{j=1}^{m} \frac{1}{f(x|\theta^{\star}_{(j)})}\Big]^{-1}, \tag{5.33}$$

mas sendo $\{\theta^{\star}_{(j)}; j = 1, \ldots, m\}$ uma amostra da distribuição *a posteriori* $h(\theta|x)$. Para se justificar este resultado, dado que se admite que $h(\theta)$ é uma distribuição própria, basta ter em conta o teorema de Bayes para se escrever

$$\frac{1}{p(x)} = \int \frac{1}{p(x)}\, h(\theta)\, d\theta = \int \frac{1}{f(x|\theta)}\, h(\theta|x)\, d\theta$$

donde decorre a estimativa de $p(x)$ expressa pela média harmónica indicada. Novamente esta estimativa tende a ser instável por causa da pequena magnitude de certos valores de $f(x|\theta)$.

Gelfand e Dey (1994) sugeriram uma generalização de (5.33) que dá origem a uma estimativa mais estável. Para tal há necessidade de considerar uma distribuição própria $g(\theta)$ (por exemplo, uma distribuição Normal multivariada com vector média e matriz de covariância estimadas a partir da amostra simulada da distribuição *a posteriori*), que aproxime bem a distribuição *a posteriori* $h(\theta|x)$ e que seja fácil simular dela. Assim tem-se

$$\frac{1}{p(x)} = \int \frac{1}{p(x)}\ g(\theta)\ d\theta = \int \frac{g(\theta)}{f(x|\theta)\ h(\theta)}\ h(\theta|x)\ d\theta,$$

e consequentemente, se $\{\theta^{\star}_{(j)}, j = 1, \ldots, m\}$ for uma amostra da distribuição *a posteriori* $h(\theta|x)$, uma estimativa para $p(x)$ é

$$\hat{p}(x) = \left[\frac{1}{m}\sum_{j=1}^{m} \frac{g(\theta^{\star}_{(j)})}{f(x|\theta^{\star}_{(j)})\ h(\theta^{\star}_{(j)})}\right]^{-1}. \tag{5.34}$$

Informações adicionais sobre estas e outras estimativas da densidade preditiva *a priori* podem encontrar-se designadamente em Kass e Raftery (1995).

5.3.3 Amostragem das distribuições preditivas

Suponha-se que se quer amostrar da distribuição preditiva $p(y|x)$, onde $y = (y_1, \ldots, y_{n^*})$. Se para cada elemento $\theta^{\star}_{(j)}$ da amostra $\{\theta^{\star}_{(j)}\}_{j=1}^{m}$ de $h(\theta|x)$ se simular um conjunto de dados $y^{\star}_{(j)}$ de $f(y|\theta^{\star}_{(j)})$, então marginalmente $y^{\star}_{(j)}$ é uma amostra de $p(y|x)$. Consequentemente $y^{\star}_{r,(j)}$, o r-ésimo elemento da amostra $y^{\star}_{(j)}$, é uma observação de $p(y_r|x)$. Este modo de amostrar é geralmente suficiente para investigar a adequabilidade do modelo como é proposto pelos resíduos bayesianos padronizados d_r.

Mas como se poderá amostrar da distribuição preditiva condicional denotada por $p(y_i|x_{(-i)})$? Seguindo o mesmo raciocínio, bastará para tal obter uma amostra $\{\theta^{\star\star}_{(j)}\}_{j=1}^{m}$ de $h(\theta|x_{(-i)})$ e para cada $\theta^{\star\star}_{(j)}$, obter uma amostra de $f(y_i|\theta^{\star\star}_{(j)})$. Obviamente que este procedimento é fastidioso e nada eficiente em amostras grandes.

O problema que se põe aqui é de como amostrar de $h(\theta|x_{(-i)})$ para todo o i, sem ter de repetir o ciclo respeitante à retirada sucessiva de cada observação, isto é, à custa da amostra $\{\theta^{\star}_{(j)}\}_{j=1}^{m}$ de $h(\theta|x)$. Note-se que para cada $\theta^{\star}_{(j)}$ se

tem para $x = (x_i, x_{(-i)})$

$$h(\theta^{\star}_{(j)}|x_{(-i)}) = \frac{p(x_i|x_{(-i)})}{f(x_i|x_{(-i)}, \theta^{\star}_{(j)})}\ h(\theta^{\star}_{(j)}|x) \propto \frac{1}{f(x_i|x_{(-i)}, \theta^{\star}_{(j)})}\ h(\theta^{\star}_{(j)}|x).$$

Assim, se se reamostrar de $\{\theta^{\star}_{(j)}\}_{j=1}^{m}$, com probabilidades proporcionais a $\omega_j = \{f(x_i|x_{(-i)}, \theta^{\star}_{(j)})\}^{-1}$ e com reposição, a amostra resultante é aproximadamente de $h(\theta|x_{(-i)})$. Muito frequentemente ocorre que $h(\theta|x_{(-i)}) \approx h(\theta|x)$ e, portanto, a reamostragem é desnecessária.

Para amostrar da distribuição marginal $p(x)$, desde que ela seja própria, basta amostrar $\tilde{\theta}_j$ de $h(\theta)$ e depois obter uma amostra $\tilde{x}_j$ de $f(x|\tilde{\theta}_j)$.

Capítulo 6

Métodos de Monte Carlo em Cadeias de Markov

Como se referiu no Cap. 4, a obtenção de inferências bayesianas na generalidade dos problemas requer uma via empírica assente numa amostra de valores simulados da distribuição *a posteriori*, tipicamente multivariada, diga-se $h(\theta|x)$, $\theta \in \Theta$, como forma de fazer face à inviabilização da via analítica exata. Dependendo da complexidade desta distribuição-alvo, o cálculo de quantidades *a posteriori* como $E[g(\theta)|x]$ pode ser realizado pelos métodos de Monte Carlo (MC) clássicos, através da concretização de amostras i.i.d. daquela distribuição geradas diretamente dela ou de uma apropriada distribuição instrumental que a envolva.

Em situações de maior complexidade começou-se a recorrer, particularmente a partir da última década do século XX, a métodos de MC mais abrangentes baseados na realização de uma cadeia de Markov homogénea convergente para a distribuição-alvo $\pi(\theta) \equiv h(\theta|x)$. Estes métodos designados por **métodos de Monte Carlo em Cadeias de Markov** (MCMC) apoiam-se assim em amostras de θ dependentes, o que implica resultados assintóticos mais complexos e maior número de iterações para a aplicação destes do que no caso dos métodos MC tradicionais.

A redescoberta por estatísticos nos recentes anos 90 (em lugar destaca-

díssimo, Gelfand e Smith, 1990) deste tipo de métodos[15], fez progredir consideravelmente a inferência baseada em simulação e, em especial, a análise bayesiana de modelos demasiadamente complexos para poderem ser tratados pelas vias antecedentes.

Dada a natureza dos métodos MCMC, não é possível compreender plenamente a sua essência e as particularidades da sua aplicação em Estatística Bayesiana sem algum conhecimento de alguns dos conceitos e resultados básicos de Cadeias de Markov que, por isso mesmo, são descritos na 1ª secção, aqui e ali assinalados a negrito, ainda que de forma bastante sintética devido ao âmbito desta obra[16]. Por motivos de simplicidade adota-se nessa secção (e na essência das restantes) uma notação genérica para as componentes da cadeia. Nas seguintes descrevem-se os métodos iterativos mais utilizados associados com os algoritmos de Metropolis-Hastings, de Gibbs e em fatias. A última secção é dedicada a questões associadas com a execução dos métodos MCMC, especificamente a geração de uma cadeia única *versus* cadeias múltiplas e métodos de diagnóstico para avaliação da convergência.

6.1 Noções e resultados básicos sobre cadeias de Markov

Um **processo estocástico** é qualquer coleção de variáveis aleatórias definidas sobre o mesmo espaço de probabilidade, $\{U(t), t \in T\}$, onde T é um subconjunto de $\mathbb{R}$ que, por comodidade, é entendido como uma classe de instantes de tempo. Quando esta classe é o conjunto discreto de inteiros positivos $T = \{0, 1, 2, \ldots\}$, o processo estocástico é usualmente denotado por $\{U_n, n \geq 0\}$, sendo esta a situação típica no contexto de um esquema de simulação estocástica. O conjunto $\mathcal{U}$ de valores das variáveis é denominado **espaço de estados**.

Num processo o conhecimento de estados do passado e do presente influencia geralmente a plausibilidade de ocorrência dos estados futuros. Quando fixado o estado do presente os estados do passado deixam de ter influência no

[15] Os seus precursores encontram-se em Metropolis *et al.* (1953), Hastings (1970) e Geman e Geman (1984).

[16] Para mais informação sobre este assunto, veja-se Ross (2014) e Tierney (1996), com incursão por referências neles contidas, se necessário.

futuro, diz-se que o processo goza da propriedade de **dependência de Markov**. O processo $\{U_n, n \geq 0\}$ satisfazendo tal propriedade de independência condicional é denominado **cadeia de Markov**, podendo ser definido através de

$$U_{n+1} \perp\!\!\!\perp (U_0, \ldots, U_{n-1}) | U_n \Leftrightarrow$$
$$P(U_{n+1} \in A | U_0 = u_0, \ldots, U_n = u) = P(U_{n+1} \in A | U_n = u) \equiv P_n(u, A),$$

para todo o acontecimento A e $n \geq 0$, onde o símbolo $P_n(u, A)$ denota a chamada **função de transição** (em um passo) quando parte do instante n. Equivalentemente, tomando $A =]-\infty, v]$ a cadeia de Markov pode ser definida através das funções de distribuição condicionais por

$$F_{U_{n+1}}(v | U_0 = u_0, \ldots, U_n = u) = F_{U_{n+1}}(v | U_n = u) \equiv F_n(u, v),$$

para todo o $v, u \in \mathcal{U}$. Quando a função de transição é invariável com n, sendo então denotada por $P(u, A)$ (ou $F(u, v)$), a cadeia de Markov diz-se **homogénea**. No quadro usual que se vai considerar, só nos interessarão as cadeias homogéneas, pelo que este qualificativo será doravante omitido.

Quando o espaço de estados é um conjunto discreto, a cadeia de Markov (implicitamente homogénea, como se alertou) fica cabalmente definida através das funções de probabilidade condicionais $P(u, \{v\})$, *i.e.*

$$P(U_{n+1} = v | U_0 = u_0, \ldots, U_n = u) = P(U_{n+1} = v | U_n = u) \equiv p(u, v), \forall n \geq 0, u, v \in \mathcal{U}.$$

A função de transição $p(\cdot, \cdot)$ é apresentável na forma de uma matriz P de probabilidades (de transição em um passo) no caso de um número finito de estados. Quando $\mathcal{U}$ é infinito não numerável e $F(u, v)$ é absolutamente contínua, a função de transição pode passar a ser definida pela correspondente densidade $p(u, v) = \frac{\partial F(u,v)}{\partial v}$.

Restringindo-nos, a menos que se explicite o contrário, a uma cadeia de Markov com espaço de estados discreto, tem-se

$$P(U_{n+1} = v) = \sum_u P(U_n = u) p(u, v) = \sum_u P(U_0 = u) p^n(u, v),$$

onde $p^n(u, v) = P(U_n = v | U_0 = u) = \sum_u p^{n-1}(u, z) p(z, v)$, $n \geq 1$ traduz a função de transição em n passos (cuja forma matricial é definida pelo produto P^n, atendendo à sua definição). A construção de uma cadeia de Markov é

assim completamente determinada pela sua função de transição, desde que se conheça a distribuição inicial.

Diz-se que uma distribuição de probabilidade $\pi(u), u \in \mathcal{U}$ é **estacionária** se

$$\pi(v) = \sum_u \pi(u)p(u,v).$$

Em particular, a distribuição inicial $P(U_0 = u) = \pi(u)$ é estacionária sse a distribuição de U_n é invariante com n, *i.e.* $P(U_n = u) = \pi(u), \forall n \geq 0$.

A existência e unicidade de distribuições estacionárias depende de a cadeia possuir algumas propriedades de estabilidade conhecidas como irredutibilidade e recorrência. Uma cadeia é **irredutível** se pode atingir qualquer estado quando parte dele ou de outro estado qualquer em um número finito de transições. Diz-se **recorrente** se volta infinitas vezes ao estado de onde parte, qualquer que ele seja e, em particular, diz-se **recorrente positiva** se é finito o tempo médio de retorno (número médio de passos) até u quando parte dele, $\forall u$. Note-se que a irredutibilidade implica recorrência positiva se $\mathcal{U}$ é finito.

Toda a cadeia de Markov (com $\mathcal{U}$ discreto) irredutível e recorrente positiva apresenta uma distribuição estacionária única. Por outro lado, existindo uma distribuição estacionária $\pi(v)$ tal que $\lim_{n\to\infty} p^n(u,v) = \pi(v)$, então a distribuição estacionária é única satisfazendo $\lim_{n\to\infty} P(U_n = v) = \pi(v)$. Sendo assim, independentemente da distribuição inicial, para valores suficientemente grandes de n a distribuição de U_n é aproximadamente dada por π.

A convergência para a distribuição estacionária π não é garantida por uma cadeia irredutível e recorrente positiva. Todavia, se se impuser a essa cadeia a condição adicional de ser **aperiódica**, traduzida por $min\{n \geq 1 : p^n(u,u) > 0\} = 1$ - basta exigir que $\exists u, p(u,u) > 0$ -, tal cadeia denominada então **ergódica** apresenta o comportamento limite $p^n(u,v) \underset{n\to\infty}{\longrightarrow} \pi(v), \forall u, v \in \mathcal{U}$, assegurando assim a convergência de $P(U_n = u)$ para $\pi(u), \forall u$.

Adicionalmente, se $g(U)$ for uma variável aleatória função daquela associada com a cadeia de Markov ergódica, com valor esperado finito segundo π, tem-se que

$$\frac{1}{n}\sum_{t=1}^{n} g(U_t) \xrightarrow[n\to\infty]{qc} E_\pi\left[g(U)\right].$$

Este resultado, conhecido usualmente como **Teorema Ergódico**, gene-

raliza a Lei Forte dos Grandes Números para cadeias markovianas com as caraterísticas mencionadas. Sob condições adicionais, também existem extensões do Teorema Limite Central refletindo a convergência em distribuição para uma distribuição Normal da sequência $\sqrt{n}\left[\frac{1}{n}\sum_{t=1}^{n} g(U_t) - E_\pi\left[g(U)\right]\right]$.

Quando as componentes da cadeia são variáveis aleatórias absolutamente contínuas, a definição de propriedades descritivas do comportamento da cadeia deve apelar a acontecimentos de $\mathcal{U}$ em substituição dos estados tomados individualmente, na linha do que se fez com a função de transição, e a não ignorar as tecnicidades próprias de tais medidas de probabilidade. Por exemplo, a descrição da dinâmica da cadeia quando parte de cada estado envolve as visitas a acontecimentos A de probabilidade positiva. A medida de probabilidade Π diz-se então estacionária se para todo o acontecimento A

$$\Pi(A) = \int_{\mathcal{U}} P(u, A)\Pi(du),$$

o que corresponde essencialmente, em termos das respetivas densidades, a

$$\pi(v) = \int_{\mathcal{U}} p(u, v)\pi(u)du.$$

Os resultados de convergência para cadeias com espaço de estados infinito não numerável são análogos aos referidos atrás, com a diferença de exigirem requisitos mais fortes[17].

Uma outra propriedade de cadeias de Markov, de importância teórica e prática na análise do seu comportamento limite, diz respeito à imutabilidade da sua dinâmica probabilística com a direção progressiva ou regressiva com que ela é analisada. Em termos concretos, uma cadeia diz-se **reversível** se para todo o acontecimento A e u do espaço de estados $\mathcal{U}$ (discreto ou não)

$$P(U_{n+1} \in A | U_n = u) = P(U_{n+1} \in A | U_{n+2} = u), \forall n.$$

A reversibilidade de uma cadeia com função de transição $p(\cdot, \cdot)$ e distribuição estacionária $\pi(\cdot)$ equivale à verificação da relação

$$\pi(u)p(u, v) = \pi(v)p(v, u), \ \forall u, v \in \mathcal{U},$$

conhecida como **condição detalhada de equilíbrio** por exprimir um equilíbrio no fluxo da cadeia no sentido de os movimentos de estar em u e passar

[17] Como a recorrência positiva à Harris para o conceito de ergodicidade - *vide e.g.* o livro Paulino *et al.* (2018) e referências nele citadas.

a v e de estar em v e passar a u serem igualmente plausíveis para todo o par (u, v) de estados. Assim, em particular, uma cadeia que satisfaz esta condição onde π é uma função probabilidade ou densidade, não só é reversível como apresenta uma distribuição estacionária dada precisamente por π.

6.2 Algoritmo de Metropolis-Hastings

Aqui e nas secções imediatamente seguintes manter-se-á essencialmente a notação da secção anterior. Dada a natureza dominantemente multivariada dos elementos das cadeias e por motivos de estética notacional, começa-se já a mudar o índice do "tempo" para a posição superior do símbolo dos vetores aleatórios originando $U^{(t)}$, reservando a posição inferior para as suas componentes (geralmente univariadas) $U_j^{(t)}$, que interessa por vezes explicitar. Note-se que isto destoa da notação usada nos Caps. 4 e 5. Na verdade, representando U o vetor k-paramétrico θ ($k \geq 2$), então $U^{(t)}$ e $U_j^{(t)}$ identificam o que ali foi denotado por $\theta_{(t)}$ e $\theta_{(t)j}$[18]. A distribuição estacionária continua a denotar-se por $\pi(u), u \in \mathcal{U}$[19].

O instrumento fundamental deste algoritmo é uma distribuição condicional $q(v|u) \equiv q(u, v)$ à qual está reservado o papel de gerador de valores simulados (definindo uma cadeia se essa distribuição for uma correspondente função de transição). Por isso, um requisito básico de $q(\cdot|\cdot)$, por vezes etiquetada de distribuição instrumental, é que permita uma fácil simulação dela. Os valores $v^{(t)}$ dela gerados sucessivamente são sujeitos a um crivo estocástico, baseado em $q(\cdot|\cdot)$ e $\pi(\cdot)$, que determina a aceitação ou rejeição de cada um deles, em que o valor substituindo o $v^{(t)}$ rejeitado é o anterior valor simulado que foi aceite. Este procedimento é descrito especificamente pelo seguinte algoritmo:

Algoritmo de Metropolis-Hastings (M-H)

1. Dado $u^{(t)}$, $t = 0, 1, 2, \ldots$ gere-se um valor de $V^{(t)} \sim q(v|u^{(t)})$.
2. Calcule-se o valor do rácio M-H $R(u^{(t)}, V^{(t)})$, em que $R(u, v) = \frac{\pi(v)q(u|v)}{\pi(u)q(v|u)}$,

[18] Alerta-se que, para $k = 1$, $U^{(t)}$ corresponde a θ_t nesses capítulos precedentes, sendo aí reservado $\theta_{(t)}$ para o t-ésimo menor valor de uma amostra $(\theta_1, \ldots, \theta_n)$.

[19] Por comodidade de uso terminológico, adota-se o termo função "densidade" para representar distribuições independentemente da natureza do seu suporte, correspondendo pois a probabilidades pontuais no caso de variáveis discretas.

e considere-se a probabilidade $\alpha(u,v) = min\{R(u,v),1\}$.

3. Tome-se o próximo valor da cadeia como a concretização de

$$U^{(t+1)} = \begin{cases} V^{(t)}, & \text{com probabilidade } \alpha(u^{(t)}, V^{(t)}) \\ u^{(t)}, & \text{com probabilidade } 1-\alpha(u^{(t)}, V^{(t)}). \end{cases} \tag{6.1}$$

Um conjunto de observações pertinentes sobre este algoritmo expõem-se a seguir separadamente num formato de notas.

Nota 1: A probabilidade de aceitação de $v^{(t)}$ na iteração $t+1$ requer que $\pi(u^{(t)}) > 0$. Esta condição é garantida $\forall t \in I\!N$ se o valor inicial $u^{(0)}$ da cadeia a satisfaz já que todos os valores simulados com $\pi(v^{(t)}) = 0$ são rejeitados pelo facto de $\alpha(u^{(t)}, v^{(t)}) = 0$. Convenciona-se que $R(u,v) = 0$ na situação em que $\pi(v) = 0 = \pi(u)$. Assim, logo que entre no suporte de π a cadeia não sai dele (q.c.).

Nota 2: A forma da razão M-H mostra que a implementação deste algoritmo é possível se $\pi(\cdot)$ e $q(\cdot|u)$ forem conhecidas a menos de constantes normalizadoras, entendidas como independentes de u no caso de $q(\cdot|u)$. Por outro lado, os valores $v^{(t)}$ para os quais a razão $\pi(v^{(t)})/q(v^{(t)}|u^{(t)})$ é superior à do seu antecessor, $\pi(u^{(t)})/q(u^{(t)}|v^{(t)})$, são sempre aceites pois $\alpha(u^{(t)}, v^{(t)}) = 1$.

Nota 3: A cadeia $\{u^{(t)}\}$ construída através deste algoritmo pode incluir repetições de alguns dos valores distintos e é uma concretização de uma cadeia de Markov já que a distribuição de $U^{(t+1)}$ condicionalmente a todas as observações anteriores depende apenas do valor de $U^{(t)}$. A convergência desta cadeia para a distribuição-alvo $\pi(u)$ está expectavelmente dependente da distribuição instrumental.

Nota 4: Dada a sua relevância nas aplicações, considera-se aqui o caso absolutamente contínuo (em relação à medida de Lebesgue) de um conjunto infinito não numerável de estados, no qual $\pi(u)$ representa uma densidade da distribuição estacionária.

Denote-se por $Q(\cdot,\cdot)$ a função de transição de uma cadeia de Markov de densidade $q(\cdot,\cdot)$, *i.e.* $Q(u,dv) = q(u,v)dv$. Atendendo ao passo 3 do algoritmo M-H, a função de transição da respetiva cadeia pode ser definida por

$$P(u,dv) \equiv P\left[U^{(t+1)} \in dv | U^{(t)} = u\right] = \alpha(u,v)q(u,v)dv + r(u)\delta_u(dv), \tag{6.2}$$

onde $\delta_u(dv)$ denota a medida de Dirac em dv e $r(u) = 1 - \int \alpha(u,v)q(u,v)dv$

representa a probabilidade de o algoritmo permanecer em u. Na situação-limite tem-se a "densidade" de transição

$$p(u,v) = \alpha(u,v)q(u,v) + r(u)\delta_u(v),$$

a qual, por definição de α e δ_u, verifica a condição detalhada de equilíbrio com π, $\pi(u)p(u,v) = \pi(v)p(v,u)$. Consequentemente, a cadeia M-H é reversível tendo como uma distribuição estacionária precisamente a distribuição pretendida π.

Nota 5: A convergência da cadeia de Markov M-H para a distribuição estacionária π depende da verificação de propriedades de estabilidade à luz do que se referiu na secção anterior.

Sendo $\mathcal{S} = \{u : \pi(u) > 0\}$ o suporte de π, o uso de uma função de transição $q(\cdot|\cdot)$ satisfazendo $q(v|u) > 0$, $\forall (u,v) \in \mathcal{S} \times \mathcal{S}$ assegura que a cadeia $\{U^{(t)}\}$ é irredutível relativamente a π. Como π é uma distribuição estacionária da cadeia M-H, esta é recorrente positiva (e também recorrente à Harris), podendo aplicar-se o disposto no teorema ergódico mencionado na Secção 6.1[20].

Em suma, numa cadeia M-H convergente para a distribuição-alvo π verifica-se que a partir de um certo ponto os valores da cadeia podem ser vistos aproximadamente como simulações de π quando na realidade foram gerados da distribuição instrumental, implicando que caraterísticas daquela distribuição possam ser determinadas empiricamente duma amostra da cadeia. ■

O traço talvez mais apelativo do algoritmo M-H é a sua versatilidade e abrangência tendo em consideração os requisitos limitados exigidos às distribuições π e q para se garantir a convergência da cadeia para π. Deve-se desde já alertar que a ocorrência dessa convergência não significa por si só um bom funcionamento do algoritmo no sentido de se atingir o equilíbrio da cadeia de um modo relativamente rápido e com uma boa representação na cadeia do conjunto de estados.

Uma distribuição instrumental bem escolhida deve produzir valores simulados que cobrem o suporte da distribuição-alvo num bom número de iterações e que não sejam aceites ou rejeitados muito frequentemente, aspetos estes que

[20] *Vide e.g.* Tierney (1994) ou Robert e Casella (2004). Se a cadeia for adicionalmente aperiódica - e isto é garantido se $r(u) > 0, \forall u \in \mathcal{S}$ -, pode provar-se a convergência (num determinado sentido) da função de transição em n passos, $P^n(u,\cdot)$, para a medida Π de que π é densidade quando $n \to \infty$.

estão relacionados com a dispersão da distribuição proposta para gerar as simulações. Com efeito, se q for demasiado dispersa em relação a π, os valores dela reproduzidos são rejeitados frequentemente e o suporte de π só poderá ser representativamente amostrado após muitas iterações, configurando uma convergência demasiado lenta. Na situação oposta de uma pequena dispersão, apenas uma região de $\mathcal{S}$ será visitada em muitas iterações na sequência de uma alta taxa de aceitação, podendo dar a entender enganadoramente uma convergência rápida, quando outras regiões por serem indevidamente exploradas vão exigir mais e mais iterações até se alcançar uma boa cobertura do suporte-alvo. Por tudo isto deve-se tentar fazer uma análise preliminar da parte conhecida de π de modo a escolher uma fonte q de simulações que reproduza a distribuição-alvo o melhor possível.

Dado o caráter genérico do algoritmo M-H, descreve-se agora com propósitos dominantemente ilustrativos dois dos seus casos particulares[21].

Algoritmo M-H com independência

A designação deste algoritmo significa tão-só que a distribuição instrumental não depende de modo algum das iterações, *i.e.* $q(v|u) = q(v)$. Isto implica que a probabilidade de aceitação de cada valor dela gerado seja

$$\alpha(u^{(t)}, v^{(t)}) = min\left\{\frac{\pi(v^{(t)})q(u^{(t)})}{\pi(u^{(t)})q(v^{(t)})}, 1\right\},\ t \geq 0,$$

variando assim com o valor aceite da iteração anterior. De acordo com o que se disse para o algoritmo M-H, a ergodicidade da cadeia $\{U^{(t)}\}$ depende da condição de o suporte da distribuição instrumental q, agora não condicionada, incluir o de π.

Uma ilustração deste algoritmo enquadrado na simulação de uma distribuição *a posteriori*, *i.e.* $\pi(\theta) = h(\theta|x) \propto L(\theta|x)h(\theta)$, em que $\{U^{(t)} \equiv \theta^{(t)}\}$, é concretizada quando se considera $q(\theta) = h(\theta)$ - note-se que neste caso o suporte de q cobre o de π, ainda que as duas densidades possam ser bem distintas. A razão M-H neste caso particulariza-se numa razão de verosimilhanças, $R(\theta^{(t)}, V^{(t)}) = \frac{L(V^{(t)}|x)}{L(\theta^{(t)}|x)}$.

Algoritmo M-H com passeio aleatório

Este algoritmo parte de uma cadeia de Markov na simulação da distribui-

[21] Veja-se Givens e Hoeting (2005) para outros casos.

ção instrumental definida por $V^{(t)} = U^{(t)} + \varepsilon_t$, onde ε_t representa um erro aleatório com uma distribuição q^* independente de $U^{(t)}$. Trata-se assim de um passeio aleatório associado com a densidade de transição $q(v|u) = q^*(v-u)$. Escolhas usuais para q^* incluem distribuições uniformes sobre uma bola centrada na origem, gaussianas e t-Student.

Note-se que se a distribuição instrumental for simétrica, i.e. $q(v|u) = q(u|v)$, as razões M-H simplificam-se para $R(u,v) = \frac{\pi(v)}{\pi(u)}$, evidenciando bem que elas dispensam o conhecimento da constante normalizadora da distribuição-alvo. No caso vertente, a simetria ocorre se $q^*(y)$ depender de y através de $|y|$. Quando a cadeia gerada pelo passeio aleatório for obtida de $V^{(t)} \sim q^*(|v - u^{(t)})|)$ cai-se num esquema do algoritmo de Metropolis[22], apresentado em Metropolis *et al.* (1953) no contexto de um problema da Física das Partículas com um espaço de estados discreto.

6.3 Amostrador de Gibbs

O caráter genérico do algoritmo M-H ficou bem patente na sua descrição na secção precedente ao ponto de não ser necessário, nomeadamente, discriminar a dimensionalidade da distribuição-alvo $\pi(u)$. Em contrapartida, o algoritmo de amostragem Gibbs[23], que vai ser o objeto desta secção, é especialmente talhado para situações k-variadas ($k \geq 2$).

Com efeito, a construção da cadeia de Markov que se pretende que convirja para a distribuição $\pi(u)$, $u = (u_1, u_2, \ldots, u_k) \in \mathcal{U}$ é nele feita sequencialmente por amostragem das distribuições condicionais (tipicamente univariadas) dado todas as outras componentes, por isso apelidadas de distribuições condicionais completas. O algoritmo vai sucessivamente substituindo num ciclo de k passos as componentes do vetor corrente u de modo que no passo j apenas u_j é trocado pelo valor amostrado da densidade condicional $\pi_j(v_j|\{u_i, i \notin j\}), j = 1, 2, \ldots, k$. Por exemplo, tomando $k = 3$, um ciclo de 3 passos visando substituir o valor corrente u é representável pelo esquema

$$(u_1, u_2, u_3) \xrightarrow{it.1} (v_1, u_2, u_3) \xrightarrow{it.2} (u_1, v_2, u_3) \xrightarrow{it.3} (u_1, u_2, v_3).$$

[22] Nicholas Metropolis, juntamente com Stanislaw Ulam, foram os pais do que eles cunharam como métodos de Monte Carlo.

[23] A designação deste método vem de Geman e Geman (1984) na sua aplicação ao estudo dos denominados campos aleatórios de Gibbs por alusão ao físico J. W. Gibbs.

Para a descrição formal do amostrador Gibbs, seja $U = (U_1, U_2, \ldots, U_k)$ o vetor aleatório com densidade $\pi(u)$ e denote-se por U_{-j} o vetor U sem a j-ésima componente, $U_{-j} = (U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_k)$, $j = 1, 2, \ldots, k$.

Amostrador Gibbs básico

1. Dado $u^{(t)} = (u_1^{(t)}, \ldots, u_k^{(t)})$, iniciado em $t = 0$, amostre-se cada uma das componentes, $v_j^{(t)}$, do próximo vetor da cadeia a partir da distribuição $V_j^{(t)} \sim \pi_j(v_j^{(t)} | u_{-j}^{(t)})$, para $j = 1, 2, \ldots, k$.

2. Findo o ciclo de k iterações, tome-se $u^{(t+1)} = (v_1^{(t)}, \ldots, v_k^{(t)})$ e repita-se o ciclo 1. substituindo t por $t + 1$.

Tudo se passa como se no j-ésimo passo Gibbs do ciclo assente em $u^{(t)}$, para todo o j, se obtivesse um vetor $v^{(t)} = (u_1^{(t)}, \ldots, u_{j-1}^{(t)}, v_j^{(t)}, u_{j+1}^{(t)}, \ldots, u_k^{(t)})$ tal que

$$V^{(t)} | u^{(t)} \sim q_j(v^{(t)} | u^{(t)}) = \begin{cases} \pi_j(v_j^{(t)} | u_{-j}^{(t)}), & \text{se } v_{-j}^{(t)} = u_{-j}^{(t)} \\ 0, & \text{c.c. .} \end{cases} \tag{6.3}$$

O facto de nesta versão dita básica do algoritmo Gibbs só ser simulado em cada passo uma componente de U, sem aproveitamento imediato no passo seguinte da atualização nos passos anteriores, tende a retardar (ou mesmo a inviabilizar em alguns casos) o processo de convergência. A designação amostrador Gibbs não se confina a esta versão básica, antes englobando variantes que incidem sobre o modo dos processos de atualização sequencial e de simulação. Entre elas, destacam-se as seguintes:

- **Algoritmo Gibbs padrão (com pronta atualização)**

  Este algoritmo adota um processo de simulação por passos em cada ciclo que se baseia no uso imediato dos valores resultantes da atualização sequencial nesse ciclo das componentes precedentes. Isto é, dado $u^{(t)} = (u_j^{(t)}, j = 1, \ldots, k)$ a simulação de $U_j^{(t+1)}$ é feita da respetiva distribuição condicional dado $u_m^{(t+1)}$, $1 \leq m < j$ e $u_m^{(t)}$, $j + 1 \leq m \leq k$.

  É compreensível que esta variante seja mais eficaz do que a versão básica no que concerne à velocidade e minúcia com que a cadeia gerada percorre o suporte da distribuição-alvo, sendo por isso considerada a versão padrão do algoritmo Gibbs. Eis a sua descrição concreta:

Dado $u^{(t)} = (u_1^{(t)}, \ldots, u_k^{(t)})$, iniciado em $t = 0$, amostre-se sucessivamente os valores $u^{(t+1)}$ do próximo vetor da cadeia a partir das distribuições condicionais completas

$$\begin{aligned}
U_1^{(t+1)} &\sim \pi_1(u_1|u_2^{(t)}, u_3^{(t)}, \ldots, u_k^{(t)}) \\
U_2^{(t+1)} &\sim \pi_2(u_2|u_1^{(t+1)}, \ldots, u_k^{(t)}) \\
&\downarrow \\
U_{k-1}^{(t+1)} &\sim \pi_{k-1}(u_{k-1}|u_1^{(t+1)}, u_2^{(t+1)}, \ldots, u_{k-2}^{(t+1)}, u_k^{(t)}) \\
U_k^{(t+1)} &\sim \pi_k(u_k|u_1^{(t+1)}, u_2^{(t+1)}, \ldots, u_{k-1}^{(t+1)}),
\end{aligned} \tag{6.4}$$

e repita-se o ciclo de k passos a partir de $u^{(t+1)}$.

Exemplo 6.1 Seja $x = (x_i, i = 1, \ldots, n)$ uma concretização de uma amostra aleatória do modelo Weibull com parâmetros de escala e de forma denotados por δ e α, respetivamente, cuja função de verosimilhança é

$$L(\delta, \alpha|x) = (\delta\alpha)^n \Big(\prod_{i=1}^n x_i\Big)^{\alpha-1} e^{-\delta \sum_i x_i^\alpha}, \ \delta, \alpha > 0.$$

Admita-se que *a priori* δ e α são independentes com distribuições Gama, $Ga(a, b)$, e Log-normal, $LN(c, d)$, de hiperparâmetros completamente especificados, com $a, b, d > 0$ e $c \in \mathbb{R}$.

A densidade conjunta *a posteriori* apresenta um núcleo dado por

$$h(\delta, \alpha|x) \propto \alpha^{n+c/d-1} \Big(\prod_{i=1}^n x_i\Big)^\alpha e^{-(\ln\alpha)^2/2d} \ \delta^{a+n} e^{-\delta\left(b+\sum_i x_i^\alpha\right)}.$$

A inspeção desta função para efeitos de identificação das distribuições condicionais completas mostra que

i) $h(\delta|\alpha, x) \propto \delta^{a+n} e^{-\delta\left(b+\sum_i x_i^\alpha\right)}$;

ii) $h(\alpha|\delta, x) \propto \alpha^{n+c/d-1} \Big(\prod_{i=1}^n x_i\Big)^\alpha e^{-\left[\frac{(\ln\alpha)^2}{2d} + \delta \sum_i x_i^\alpha\right]}$.

A distribuição condicional completa de δ é assim $Ga(a + n, b + \sum_i x_i^\alpha)$ pelo que a geração de valores de δ para cada α é factível diretamente através de disponíveis algoritmos eficientes de simulação da distribuição Gama. A distribuição condicional completa de α já não apresenta uma

estrutura padrão exigindo por isso o recurso a métodos mais sofisticados de simulação estocástica. Um deles é o método de rejeição adaptativa dada a concavidade da densidade Log-normal, incluído na classe referida na Nota 4 desta secção.

■

- **Algoritmo Gibbs com agrupamento**

  Embora a descrição típica do algoritmo Gibbs faça uso das distribuições condicionais completas univariadas, o esquema de simulação pode apresentar uma estrutura mais flexível com um número de passos inferior à dimensionalidade da distribuição-alvo. A variante em questão tem a particularidade de reunir em grupos ou blocos as variáveis mais intercorrelacionadas de modo que a simulação conjunta a partir das distribuições condicionais completas dos blocos possa acelerar a convergência do algoritmo, como se tem constatado em modelos de grande complexidade paramétrica.

- **Algoritmo Gibbs com hibridação**

  Em geral, o conjunto das distribuições condicionais completas requerem a escolha e aplicação de métodos diferenciados de simulação estocástica. Frequentemente surgem nele distribuições para as quais se desconhece que meios de simulação utilizar. Nesses passos Gibbs há sempre a possibilidade de se recorrer a outros algoritmos iterativos, hibridizando assim o amostrador de Gibbs. É o que acontece com o uso dentro deste de uma ou outra variante do algoritmo M-H de que são exemplo os algoritmos propostos por Müller (1991,1993).

Feita uma rápida menção a algumas das suas variantes, tem interesse destacar alguns aspetos do amostrador Gibbs em geral, o que será feito nas notas seguintes.

Nota 1: Pelo exposto, as fontes de simulação no amostrador de Gibbs advêm da própria distribuição-alvo, evitando o problema da escolha, nem sempre fácil, de uma "boa" distribuição instrumental requerida no algoritmo M-H. Mas também é verdade que a amostragem de uma só variável em cada iteração não deixa de ser um entrave a uma rápida digressão pelo suporte da distribuição-alvo.

Nota 2: Apesar das diferenças existentes entre os algoritmos Gibbs e M-H, pode estabelecer-se uma relação entre eles quando se esquadrinha cada passo do primeiro deles na sua versão básica. De facto, viu-se já que em cada passo j de um ciclo baseado em $u^{(t)}$ a simulação processa-se através da distribuição condicional $q_j(u^{(t)}, v^{(t)}) = \pi(v_j^{(t)}|u_{-j}^{(t)})$ - a desempenhar o papel de uma distribuição instrumental -, sendo $v^{(t)}$ um vector que só difere de $u^{(t)}$ na j-ésima componente, pelo que $v_{-j}^{(t)} = u_{-j}^{(t)}$. Por definição de uma distribuição conjunta e recordando a notação em curso, pode escrever-se $\pi(u^{(t)}) = \pi(u_{-j}^{(t)})\pi(u_j^{(t)}|u_{-j}^{(t)})$, em que o 1º e 2º fatores se reportam às distribuições marginal de $U_{-j}^{(t)}$ e condicional de $U_j^{(t)}$ dado $U_{-j}^{(t)}$, respetivamente, e que coincidem assim com as correspondentes distribuições de $V_{-j}^{(t)}$ e condicional de $U_j^{(t)}$ dado $V_{-j}^{(t)}$. Consequentemente,

$$\frac{\pi(v^{(t)})}{\pi(u^{(t)})} = \frac{\pi(v_j^{(t)}|u_{-j}^{(t)})}{\pi(u_j^{(t)}|v_{-j}^{(t)})} \equiv \frac{q_j(u^{(t)}, v^{(t)})}{q_j(v^{(t)}, u^{(t)})},$$

implicando para a razão M-H desse passo

$$R_j(u^{(t)}, v^{(t)}) = \frac{\pi(v^{(t)})q_j(v^{(t)}, u^{(t)})}{\pi(u^{(t)})q_j(u^{(t)}, v^{(t)})} = 1.$$

Cada ciclo Gibbs pode então ser visto como uma composição de k passos M-H com probabilidade de aceitação por passo igual a 1. Note-se que se se encarar cada ciclo do amostrador Gibbs como um algoritmo M-H particular, a correspondente probabilidade de aceitação global calculada da densidade de transição entre o início e o fim do ciclo é inferior a 1[24].

Nota 3: A definição do algoritmo Gibbs para o caso bidimensional resumida nos passos $U_1^{(t+1)} \sim \pi_1(\cdot|u_2^{(t)})$ e $U_2^{(t+1)} \sim \pi_2(\cdot|u_1^{(t+1)})$ para $t \geq 0$, mostra claramente que a sequência $\{(U_1^{(t)}, U_2^{(t)})\}$ é uma cadeia de Markov. Que cada uma das subsequências também o é, *e.g.* $U_2^{(t)}$, vê-se da respetiva densidade de transição

$$P(u_2, v_2) = \int \pi_1(w|u_2)\pi_2(v_2|w)dw,$$

que só depende do passado através do último valor de U_2. Além disso, a definição das densidades marginais e da densidade de transição associada a

[24]Basta considerar o caso $k = 2$ e a respetiva densidade de transição $P(u, v) = \pi_1(v_1|u_2)\pi_2(v_2|v_1)$, onde o símbolo π_j representa a densidade marginal da componente U_j do par (aqui condicionada na outra componente).

U_2 leva a que

$$\pi_2(v_2) = \int \pi_2(v_2|w)\pi_1(w)dw = \int \Big[\int \pi_2(v_2|w)\pi_1(w|u_2)dw \Big]\pi_2(u_2)du_2$$
$$= \int P(u_2, v_2)\pi_2(u_2)du_2$$

evidenciando que π_2 é a distribuição estacionária da subcadeia $U_2^{(t)}$.

Uma das condições básicas para a convergência da cadeia multivariada $\{U^{(t)}\}$ é a da positividade expressa na propriedade de o suporte $\mathcal{U}$ da distribuição conjunta $\pi(\cdot)$ ser o produto cartesiano dos suportes $\mathcal{U}_j$ das distribuições marginais $\pi_j(\cdot)$. Ela assegura que a cadeia é irredutível e, no caso bivariado, a irredutibilidade das subcadeias marginais também é garantida. Se adicionalmente a função de transição for absolutamente contínua com respeito à medida de Lebesgue, com densidade expressa pelo produto das densidades condicionais completas

$$p(u, v) = \pi_1(v_1|u_2, \ldots, u_k) \times \pi_2(v_2|v_1, u_3, \ldots, u_k) \times \pi_k(v_k|v_1, v_2, \ldots, v_{k-1}),$$

a cadeia é recorrente (à Harris), implicando que π é a distribuição estacionária da cadeia $\{U^{(t)}\}$ e as suas marginais são a distribuição-limite das respetivas subcadeias, com a decorrente aplicabilidade do teorema ergódico[25].

Em suma, a estrutura e a possível convergência do algoritmo Gibbs evidenciam que as distribuições condicionais completas podem ser suficientemente informativas para gerarem a distribuição conjunta que as originou. É elucidativo explorar exemplos de incompatibilidade das distribuições condicionais para perceber a relevância de condições que asseguram a convergência para a distribuição-alvo do algoritmo Gibbs - ver para o efeito Paulino *et al.* (2018) e referências aí citadas. Em especial, deve-se ficar precavido perante a identificação das distribuições condicionais completas com alheamento da constante normalizadora, uma vez que esta pode não ser finita inviabilizando a existência de uma distribuição conjunta própria, anomalia esta que nem sempre é detetável da inspeção das cadeias geradas mas que não é raro ocorrer em modelos bayesianos com distribuições *a priori* impróprias (*vide* Robert e Casella, 2004, Cap. 10, e citações aí feitas).

Nota 4: O processo de simulação das distribuições condicionais completas depende naturalmente da estrutura destas. Em casos mais simples, o recurso

[25]E também da convergência da função de transição em n passos para π se a cadeia for também aperiódica.

ao método de inversão (da função de distribuição) ou a métodos *ad hoc* eficientes para certas distribuições conhecidas permite facilmente cumprir o disposto nos correspondentes passos Gibbs. Em casos mais complicados, este desiderato ainda é realizável através de procedimentos mais sofisticados descritos nomeadamente nas referências indicadas no introito do Cap. 4 - *vide* também Paulino *et al.* (2018) para alguns deles como os métodos de rejeição-aceitação.

Em muitos problemas estatísticos a distribuição-alvo apresenta uma forma intrincada ao ponto de se desconhecer como simular em alguns passos Gibbs. Uma estratégia por vezes bem-sucedida (como em modelos para dados incompletos) é ampliar a distribuição-alvo $\pi(u)$ introduzindo variáveis adicionais não observadas Z de modo que $\pi(u)$ seja um resultado de marginalização de uma distribuição conjunta[26] $f(u,z)$ que se revele mais apropriada para executar o processo de simulação, agora envolvendo as distribuições $f(u|z)$ e $f(z|u)$.

Exemplo 6.2 Considere-se um problema envolvendo os resultados (positivo e negativo) de um teste de diagnóstico binário de uma doença numa amostra aleatória de um número N fixado de unidades. Sendo X o número de unidades com resultado positivo nessa amostra, tem-se $X|\phi \sim Bi(N,\phi)$. O diagnóstico para a grande maioria dos testes está sujeito a erros de classificação implicando que a probabilidade de o teste acusar um resultado positivo se possa expressar por $\phi = \alpha\sigma + (1-\alpha)(1-\varepsilon)$ onde $\theta = (\alpha,\sigma,\varepsilon)$ com α a prevalência da doença, σ a sensibilidade do teste (probabilidade de um resultado corretamente positivo) e ε a especificidade do teste (probabilidade de um resultado corretamente negativo).

O vetor θ é tipicamente desconhecido e integra um conjunto de parâmetros de interesse inferencial. Todavia, dada a óbvia sobreparametrização do modelo amostral indicado, não é possível inferir sobre θ (e outras funções dele, à exceção de ϕ, por exemplo) sem recorrer a informação extra obtida de dados adicionais (*e.g.*, relativos a um teste de tipo padrão de ouro) ou de informação *a priori* de peritos na espécie de diagnóstico e doença em causa. Supondo que só se tem acesso a esta última e à sua representação em termos de distribuições Beta independentes para as componentes de θ, com os hiperparâmetros

[26] A determinação desta função pela sua ligação com a densidade de interesse é por vezes denominada desmarginalização ou completamento de $\pi(u)$.

fixados, a densidade *a posteriori* apresenta a forma analiticamente intratável

$$h(\theta|x) \propto f(x|\theta) \times \alpha^{a_p-1}(1-\alpha)^{b_p-1}\sigma^{c_s-1}(1-\sigma)^{d_s-1}\varepsilon^{c_e-1}(1-\varepsilon)^{d_e-1},\ \theta \in (0,1)^3,$$

justificando a necessidade de recurso a algum método MCMC. O amostrador de Gibbs não é particularmente feliz para lidar com esta distribuição *a posteriori* uma vez que as distribuições condicionais completas são complicadas, exigindo a procura de apropriados métodos de amostragem. O uso do amostrador de Gibbs fica facilitado se se recorrer a um mecanismo de ampliação de dados tomando os novos dados como $Y = (X, Z_1, Z_2)$, onde Z_1 (resp. Z_2) são variáveis não observadas (latentes) indicando o número de unidades com resultado verdadeiramente positivo (negativo). Um modelo para Y consistente com os dados observados X é definido por

$$f(y|\theta) = f(x|\theta)f(z_1|x,\theta)f(z_2|x,\theta),$$

em que $Z_1|x,\theta \sim Bi(x, \alpha\sigma/\phi)$ e $Z_2|x,\theta \sim Bi(N-x, (1-\alpha)(1-\varepsilon)/(1-\phi))$.

Note-se que os parâmetros probabilísticos das distribuições condicionais das variáveis latentes correspondem aos denominados valores preditivos positivo $V_+ = \alpha\sigma/\phi$ e negativo $V_- = (1-\alpha)(1-\varepsilon)/(1-\phi)$. A densidade *a posteriori* relativa aos dados ampliados y é então expressável por

$$\begin{aligned} h(\theta|y) \quad \propto \quad & f(x|\phi)(V_+)^{z_1}(1-V_+)^{x-z_1}(V_-)^{z_2}(1-V_-)^{N-x-z_2} \times \\ & \alpha^{a_p-1}(1-\alpha)^{b_p-1}\sigma^{c_s-1}(1-\sigma)^{d_s-1}\varepsilon^{c_e-1}(1-\varepsilon)^{d_e-1}. \end{aligned}$$

Esta expressão simplifica-se consideravelmente num produto de três densidades Beta. Com efeito, se se efetuar uma transformação equivalente dos dados $y = (x, z_1, z_2)$ para $y^* = (m, z_1, z_2)$, onde $m = z_1 + N - x - z_2$ é o número de unidades da amostra portadores de doença, conclui-se de $f(y|\theta)$ que

$$f(y^*|\theta) = f(m|\theta)f(z_1|m,\theta)f(z_2|m,\theta),$$

tal que $M|\theta \sim Bi(N, \alpha)$, $Z_1|m,\theta \sim Bi(m, \sigma)$ e $Z_2|m,\theta \sim Bi(N-m, \varepsilon)$.

Esta fatorização da verosimilhança, tendo em conta a da densidade conjunta *a priori* e o tipo Binomial e Beta dos seus fatores, implica que as componentes de θ são *a posteriori* também independentes com as seguintes distribuições

$$\begin{aligned} \alpha|y \sim Be(A_p, B_p), \quad & A_p = a_p + m = a_p + z_1 + N - x - z_2, \\ & B_p = b_p + N - m = b_p + x - z_1 + z_2 \\ \sigma|y \sim Be(C_s, D_s), \quad & C_s = c_s + z_1,\ D_s = d_s + N - x - z_2 \\ \varepsilon|y \sim Be(C_e, D_e), \quad & C_e = c_e + z_2,\ D_e = d_e + x - z_1. \end{aligned}$$

Estas são as densidades condicionais completas para os parâmetros em face dos dados ampliados y. Como as partes z_1 e z_2 destes não são observadas, há necessidade de serem imputadas a partir dos parâmetros, o que pode ser feito através das suas distribuições amostrais condicionais na parte observada x dos dados. Assim, o algoritmo de tipo Gibbs para amostragem da densidade conjunta *a posteriori* $h(\theta, z_1, z_2|x)$ é definido em ciclos de dois passos como segue:

Algoritmo de ampliação de dados em cadeia

1. Passo de imputação: Dado $\theta^{(0)} = (\alpha^{(0)}, \sigma^{(0)}, \varepsilon^{(0)})$, calcula-se $V_+^{(0)} = V_+(\theta^{(0)})$ e $V_-^{(0)} = V_-(\theta^{(0)})$ e geram-se os valores

$$z_1^{(1)} \sim Bi(x, V_+^{(0)}),\ z_2^{(1)} \sim Bi(N - x, V_-^{(0)}).$$

2. Passo *a posteriori*: Com base em $(z_1^{(1)}, z_2^{(1)})$, geram-se da densidade de θ para os dados ampliados as iteradas $\theta^{(1)}$ tal que

$$\alpha^{(1)} \sim Be(A_p, B_p),\ \sigma^{(1)} \sim Be(C_s, D_s),\ \varepsilon^{(1)}) \sim Be(C_e, D_e).$$

Partindo de $\theta^{(1)}$ repete-se o ciclo de dois passos, e assim sucessivamente.

Deve ser realçado que este algoritmo foi realmente introduzido sem qualquer relação com o amostrador de Gibbs por Tanner e Wong (1987) sob a designação atrás indicada, onde se prova que $h(\theta|x, z_1^{(i)}, z_2^{(i)})$ converge à medida que $i \to \infty$ para $h(\theta|x)$ sob condições gerais.

■

6.4 Amostrador em fatias

Já se referiu que a complexidade da distribuição-alvo π pode inviabilizar o seu papel como fonte de simulações sem que tal impeça a avaliação da função densidade em vários pontos $u \in \mathcal{U}$. Neste cenário, uma outra estratégia para desbloquear o processo MCMC é recorrer a uma variável auxiliar Z com o propósito de tornar fácil a simulação de uma cadeia de $(U, Z) \sim f(u, z) = \pi(u)f(z|u)$ em que $\pi(u)$ é a densidade marginal de interesse associada com a distribuição conjunta f. Interessa também que Z seja escolhida de modo que a mencionada cadeia seja convergente e agilize a exploração equilibrada do suporte $\mathcal{U}$ da correspondente subcadeia e o cálculo de caraterísticas de π.

Ora escolhendo Z de modo que $Z|U = u \sim Unif([0, \pi(u)])$ tem-se que $(U, Z) \sim Unif(\mathcal{S})$ onde $\mathcal{S} = \{(u, z) : u \in \mathcal{U}, z \in [0, \pi(u)]\}$. Uma maneira de conseguir simular de π é gerar uma cadeia de Markov que apresente como distribuição estacionária precisamente a distribuição uniforme multivariada na região $\mathcal{S}$.

O amostrador em fatias é um procedimento iterativo gerador de um passeio aleatório em $\mathcal{S}$, movendo-se alternadamente em duas direções através de distribuições uniformes. No 1º passo, ao longo do eixo real através precisamente de $Z|U = u \sim Unif([0, \pi(u)])$ e no 2 º passo, ao longo da região $\mathcal{U}$ por meio de $U|Z = z \sim Unif(\mathcal{S}(z))$, com $\mathcal{S}(z) = \{u \in \mathcal{U} : \pi(u) \geq z\}$ - note-se que a densidade marginal $f(z)$ é definida à custa da medida de Lebesgue de $\mathcal{S}(z)$.

Este método quando convergente consegue assim gerar uma amostra aproximadamente de π, a partir da correspondente subcadeia, usando apenas a avaliação de π em pontos simulados de uma distribuição uniforme. Na realidade, este esquema de amostragem só precisa de conhecer π a menos da constante normalizadora[27].

Em suma, o algoritmo deste amostrador é especificado por:

Amostrador em fatias

Dado $(u^{(t)}, z^{(t)})$, iniciado com $t = 0$, simule-se

1. $z^{(t+1)} \sim Unif(\left[0, \pi(u^{(t)})\right])$;
2. $u^{(t+1)} \sim Unif(\{u : \pi(u) \geq z^{(t+1)}\})$;

e repita-se o ciclo de dois passos incrementando t de uma unidade - entenda-se aqui $\pi(u)$ como a densidade ou o seu núcleo, consoante o que for mais vantajoso.

Destacam-se em seguida algumas observações sobre este procedimento de amostragem cujo nome se atribui a Neal (1997) e que se deve ao trabalho dele publicado em Neal (2003) e de Damien *et al.* (1999).

Nota 1: Quando U é unidimensional o funcionamento do amostrador em fatias é clarificável ilustrativamente através da representação gráfica da

[27] De facto, escrevendo $\pi(u) = c\pi^*(u)$ e tomando $Z^* = Z/c$, este procedimento é equivalente a usar $(U, Z^*) \sim Unif(\{(u, z^*) : u \in \mathcal{U}, z^* \in [0, \pi^*(u)]\})$, implicando $Z^*|U = u \sim Unif([0, \pi^*(u)])$ e $U|Z^* = z^* \sim Unif(\{u : \pi^*(u) \geq z^*\})$.

densidade $\pi(u)$ com os valores de U e Z marcados respetivamente nos eixos das abcissas e ordenadas. O ponto $(u^{(t)}, \pi(u^{(t)})$ define no eixo vertical uma fatia sobre a qual vai ser gerado o valor $z^{(t+1)}$. A intersecção da reta $Z = z^{(t+1)}$ com $\pi(u)$ define o(s) ponto(s) que delimitam no eixo horizontal uma outra fatia, a região $\mathcal{S}(z^{(t+1)})$ (intervalo ou união de intervalos) onde é gerado $u^{(t+1)}$. Na prática, a maior dificuldade deste algoritmo está no passo 2 devido ao suporte da respetiva distribuição simuladora (a fatia horizontal) poder ser complicado quando $\pi(u)$ é complexa, o que vai requerer o uso de outros métodos de simulação nesse passo (e.g., métodos de rejeição). De qualquer modo, pela sua estrutura este algoritmo lida melhor com densidades-alvo multimodais do que outros algoritmos (como por exemplo o de M-H) do ponto de vista da cobertura eficiente do suporte distribucional de interesse.

Nota 2: A estrutura do amostrador em fatias para $\mathcal{U}$ unidimensional mostra que ele pode ser visto como um caso especial do amostrador de Gibbs em dois passos, associado com a ampliação da distribuição-alvo $\pi(u)$ para a distribuição $f(u, z) = \pi(u)f(z|u)$ Uniforme em $\mathcal{S}$. A subsequência $\{U^{(t)}\}$ é então uma cadeia de Markov com a densidade de transição $P(u, v) = \int f(z|u)f(v|z)dz$ e densidade estacionária $\pi(u)$.

Esta interpretação mantém-se na sua essência para distribuições-alvo multivariadas com a diferença de o número de passos ser naturalmente maior. Em consequência, as condições de convergência do amostrador em fatias podem ser vistas à luz das que se aplicam ao amostrador Gibbs[28].

6.5 Aspetos inerentes à execução dos métodos

Como se tem vindo a referir, os métodos MCMC visam gerar iterativamente realizações de uma cadeia de Markov de tal modo que a cadeia se aproxime da sua condição de equilíbrio à medida que aumenta o número de iterações. Se numa determinada iteração t a cadeia já se encontrar no (ou "próxima" do) estado de equilíbrio, então os estados gerados aí e nas iterações seguintes podem ser considerados como realizações da distribuição-alvo, que vamos aqui tomar como sendo a distribuição *a posteriori* $h(\theta|x)$.

No entanto, sucessivas realizações de uma mesma cadeia ao longo do tempo

[28] A ligação com o algoritmo Gibbs aplica-se integralmente a uma generalização do amostrador em fatias descrito, *e.g.*, em Robert e Casella (2004).

não constituem uma amostra aleatória da distribuição-alvo devido à correlação entre os vetores $\theta^{(t)}$ gerados. Esta é a grande diferença relativamente ao contexto de aplicação dos métodos MC clássicos considerados no Cap.4 para o traçado das inferências de interesse. Os instrumentos empíricos de cálculo inferencial são os mesmos, logo que aqui seja definida a amostra a reter, mas a correspondente precisão requer medidas diferentes – por exemplo, a variabilidade das médias ergódicas não pode ser dada em amostras correlacionadas pelo erro padrão do caso i.i.d.[29]. Além disso, a justificação assintótica das quantidades calculadas exige requisitos adicionais, como se foi referindo nas secções precedentes deste capítulo.

Embora seja importante conhecer as condições que asseguram a convergência dos métodos MCMC descritos anteriormente, não é menos verdade que tais resultados são insuficientes do ponto de vista da execução desses métodos, porque a verificação na prática dessas condições é por vezes problemática e, acima de tudo, porque os próprios algoritmos não nos norteiam sobre quando devem ser parados para o cálculo das inferências pretendidas.

Daí que se revele absolutamente necessário na prática recorrer a procedimentos empíricos que analisem comportamentos das cadeias de modo a captar indícios de situações próximas da convergência (para o alvo correto) ou opostas a tal, mesmo sabendo que nenhum deles é infalível. Sem tais métodos de diagnóstico, não se pode confiar minimamente em que os resultados obtidos traduzam os objetivos em mente.

Dualidade cadeia única *versus* cadeias múltiplas

Os métodos de diagnóstico são dirigidos à monitorização da convergência das cadeias para a distribuição estacionária ou de médias empíricas para a correspondente média distribucional e, muito naturalmente, variam consoante o quadro em que integram os valores simulados, se numa única longa cadeia ou se em várias cadeias paralelas mais curtas geradas independentemente. Estes dois esquemas são descritos em seguida.

Esquema de uma única cadeia

[29]Uma estratégia é usar o referido erro padrão depois de nele ser introduzido um fator corretivo que tome em conta as correlações (*vide*, e.g., Robert e Casella, 2004, Cap. 12). Uma outra, referida em seguida, consiste em encurtar a amostra tomada de sucessivos estados da cadeia de tal modo que possa passar a ser vista como uma aproximação de uma amostra aleatória observada.

Sendo $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$ o estado inicial da cadeia,

- Gera-se uma longa realização da cadeia com comprimento (número de iterações) $t = l + k^* m$, onde

  - l é o número inicial de iterações necessárias para que a cadeia atinja supostamente o estado de equilíbrio (determinado da aplicação de métodos de diagnóstico mencionados abaixo) e o seu conjunto costuma designar-se por *período de aquecimento* (*warm-up* ou *burn-in*), o qual pode ser mais ou menos longo consoante a convergência da cadeia (que se admite) for mais ou menos lenta por especificidades de cada modelo;

  - m é o número prefixado das iterações que vão ser usadas na aplicação do método de Monte Carlo e k^* é o espaçamento ao longo da sequência $\{\theta^{(t)}\}$ entre iteradas a reter visando eliminar ou reduzir substancialmente a autocorrelação entre elas (obtido da análise do gráfico das autocorrelações para diferentes intervalos entre iteradas) de modo a obter-se uma amostra que se aproxime do tipo i.i.d..

- Extrai-se da realização original da cadeia um subconjunto de dimensão m contendo as observações $\theta^{(l+k^*)}, \theta^{(l+2k^*)}, \ldots, \theta^{(l+mk^*)}$, obtendo-se deste modo uma amostra denotada agora por $\theta_{(1)}, \ldots, \theta_{(m)}$, em que $\theta_{(j)} \equiv \theta^{(l+jk^*)}$ com base na qual são calculadas as inferências de interesse, como se indicou nos caps. 4 e 5.

A questão natural que se põe é de como escolher a dimensão m da amostra, o número inicial de iterações l e o espaçamento k^* entre iterações. Como as escolhas dependem fortemente das particularidades de cada problema, não há uma regra geral. Note-se que o valor de l esté nomeadamente dependente do estado iniciante da cadeia e da velocidade de convergência. A dimensão m está subordinada à precisão que se pretende para as inferências. O espaçamento k^* depende muito da estrutura de correlação e ambiciona acelerar a convergência das médias empíricas, ainda que à custa de uma menor eficiência da estimação (a média empírica da amostra emagrecida é menos precisa do que a da amostra total pós-aquecimento - *vide* MacEachern e Berliner, 1994).

Esquema de múltiplas cadeias

- Geram-se m cadeias de t^* (usualmente, $t^* << t$) iteradas cada uma, a

partir de m valores iniciais, geralmente todos distintos e bem espalhados pelo suporte da distribuição-alvo, por conveniência.

- Toma-se a última iterada $\theta^{(t^*)}$ de cada cadeia para formar a amostra $\theta_{(1)}, \ldots, \theta_{(m)}$ – a fixação de t^*, que deve exceder o período de aquecimento, e de m obedecem às mesmas considerações que se referiram atrás.

Cada um destes esquemas possui as suas vantagens e desvantagens, o que explica as várias opções que se registam na literatura[30]. O 1º esquema permite diminuir o custo computacional e pode fazer chegar a cadeia mais perto de $h(\theta|x)$ em algoritmos que funcionam lentamente do que a via das cadeias múltiplas, nas condições do mesmo número global de iterações.

O 2º esquema permite controlar mais facilmente a convergência para $h(\theta|x)$ ao reduzir a dependência dos estados iniciais e explorar mais cabalmente o suporte-alvo. Deteta mais provavelmente do que o 1º esquema se a aparente convergência de uma cadeia não é mais do que a indicação de ela ter ficado presa em alguma região modal que não representa inteiramente a distribuição-alvo, deixando a sugestão de equacionar reparametrizações do modelo (tarefa de difícil sucesso) ou de mudar os estados iniciantes. Em suma, o esquema de múltiplas cadeias parece largamente apropriado para conduzir estudos exploratórios prévios a correr uma longa cadeia final para efeitos de obtenção das inferências cogitadas.

Diagnóstico de convergência

São vários os métodos disponíveis para diagnóstico dos dois tipos de convergência, alguns dos quais estão automaticamente incluídos em software mais específico ou mais geral de análise bayesiana. Outros podem ser acrescentados ao software por ampliação do correspondente código para análise de cada problema estatístico. Limitamo-nos a referir alguns dos métodos mais usados, em geral de forma sucinta pelo âmbito deste texto, remetendo os leitores para referências apropriadas.

O instrumento mais conhecido para monitorização da convergência para a distribuição estacionária é a representação gráfica dos valores simulados da cadeia ao longo das sucessivas iterações, conhecido como gráfico dos traços ou

[30] Esta questão é longamente discutida no trabalho de Geyer (1992) e na discussão que se lhe segue, onde são apresentados os prós e contras de cada um dos métodos.

gráfico do historial, e sua análise ao longo de várias janelas temporais para inspeção de alguma alteração no padrão em que a cadeia se mistura no suporte da distribuição *a posteriori*. Um outro instrumento consiste na sobreposição gráfica das estimativas de densidades marginais *a posteriori*, à medida que se aumenta o número de iterações usado na estimação, para detetar a ocorrência de estabilidade. Outro tipo de método consiste no uso de testes não paramétricos para averiguação da estabilização distribucional da cadeia. Por exemplo, comparação pelo teste de Kolmogorov-Smirnov de duas subamostras correspondentes a períodos não sobrepostos da amostra emagrecida, em termos de distribuições marginais univariadas.

Para monitorização da convergência das médias empíricas de quantidades escalares $g(\theta)$, dadas por $S_m = \sum_{i=1}^{m} g(\theta_{(i)})/m$ (recorde-se o Cap. 4), uma técnica possível é construir o gráfico das somas cumulativas, dadas por $D_l = \sum_{i=1}^{l} \left[g(\theta_{(i)}) - S_m \right]$, $l = 1, \ldots, m$. Cadeias que exploram rapidamente o suporte-alvo tendem a apresentar este gráfico com um aspeto irregular e geralmente concentrado em torno de 0. No caso contrário, *i.e.* cadeias que se misturam lentamente no seio do suporte, este gráfico aparece com um aspeto mais regular e com longas digressões por valores afastados de 0. Uma variante possível deste gráfico utiliza a média empírica condicional quando g é função apenas de uma parte do parâmetro, diga-se $g(\alpha)$ quando $\theta = (\alpha, \beta)$, dada então por $S_m^c = \sum_{i=1}^{m} E\left[g(\alpha) | \beta_{(i)} \right]$, desde que os valores esperados condicionais no algoritmo Gibbs sejam determináveis explicitamente.

Existem outros métodos de avaliação de convergência baseados em ideias diferentes que apareceram pela mesma altura (início dos anos 90) e que ficaram conhecidos pelo nome das seus autores. São eles os métodos de Gelman-Rubin, Raftery-Lewis, Geweke e Heidelberger-Welch, cuja descrição pode ser encontrada no estudo comparativo de Cowles e Carlin (1996). Todos eles estão implementados em pacotes disponíveis no ambiente R, conhecidos pelos acrónimos CODA (de COnvergence Diagnostic and output Analysis), desenvolvido por Best *et al.* (1996) e Plummer *et al.* (2006), e BOA (de Bayesian Output Analysis) da autoria de Smith (2007). O Cap. 8 deste texto, dedicado a software bayesiano, exemplificará o uso desses pacotes na análise por métodos MCMC.

Capítulo 7

Métodos Baseados em Aproximações Analíticas

Desde os anos 80 que a investigação se tem centrado na procura de técnicas eficientes, e na medida do possível simples, para ultrapassar os problemas técnicos de cálculo. Várias estratégias foram sugeridas, nomeadamente a aproximação da distribuição *a posteriori* por uma distribuição Normal multivariada, a abordagem de Laplace, métodos de quadratura numérica, métodos de Monte Carlo clássicos e métodos de Monte Carlo via Cadeias de Markov (MCMC).

A abordagem por MCMC tem raízes nos trabalhos de Metropolis *et al.* (1953), Hastings (1970) e Geman e Geman (1984). A sua importância para resolver problemas em Estatística Bayesiana foi primeiramente reconhecida por Gelfand e Smith (1990) e, desde essa altura, muito se tem investigado e escrito sobre o assunto. É considerada, sem qualquer dúvida, uma técnica extremamente poderosa e decerto direcionou e transformou totalmente a investigação no domínio da Estatística Bayesiana. Permitiu também que a metodologia bayesiana alcançasse uma maior credibilidade em diversas áreas de aplicação, pondo-a assim na fronteira do conhecimento.

Avanços observados nas tecnologias usadas na obtenção de dados fizeram com que cada vez mais houvesse necessidade de modelar dados com um grau de complexidade cada vez maior. São exemplos disso os modelos espaço-temporais, modelos lineares dinâmicos, modelos lineares generalizados mis-

tos, modelos aditivos generalizados, processos de Cox log-gaussianos, modelos geoaditivos, etc. Todos estes modelos se encaixam numa classe mais vasta de modelos designada por modelos gaussianos latentes (LGM). Veja-se, por exemplo, Blangiardo e Cameletti (2015). Em teoria é sempre possível implementar algoritmos MCMC aplicados a modelos gaussianos latentes. Contudo, eles vêm associados com uma vasta gama de problemas em termos de convergência e de tempo computacional.

Recentemente, Rue *et al.* (2009) desenvolveram uma abordagem analítica por Aproximações de Laplace Encaixadas e Integradas (INLA - *Integrated Nested Laplace Approximation*) que permite obter aproximações determinísticas às distribuições *a posteriori* marginais dos parâmetros dos modelos, sendo particularmente eficiente na estimação bayesiana de modelos gaussianos latentes. O método INLA apresenta duas vantagens principais sobre as técnicas MCMC. A primeira é a rapidez computacional. Usando INLA obtém-se resultados em segundos ou minutos em modelos que levam horas ou mesmo dias a correr usando algoritmos MCMC. A segunda vantagem é que o INLA trata os modelos gaussianos latentes de uma maneira unificada, permitindo assim uma maior automatização do processo inferencial, independentemente do tipo de modelo que se está a desenvolver.

Para melhor compreensão do método proposto por Rue *et al.* (2009), este capítulo inicia-se com uma revisão das principais técnicas de aproximação analíticas desenvolvidas para implementar o paradigma bayesiano, passando-se depois à apresentação da metodologia INLA e da sua implementação usando o pacote específico do R desenvolvido para o efeito.

7.1 Métodos analíticos

7.1.1 Aproximação à distribuição Normal multivariada

Um modo adequado para tentar contornar os problemas de cálculo que surgem em inferência bayesiana passa pela exploração das propriedades assintóticas da distribuição *a posteriori*. Com efeito, Walker (1969) mostrou que para grandes valores de n e sob certas condições de regularidade, a distribuição *a posteriori* de um vetor θ k-dimensional é aproximadamente Normal multivariada.

Considere-se, para o efeito, a densidade *a posteriori*, $h(\theta|x)$, escrita na

forma

$$h(\theta|x) \propto \exp\{\ln h(\theta) + \ln f(x|\theta)\}.$$

Desenvolvendo em série de Taylor de 2ª ordem os dois logaritmos do segundo termo da expressão anterior, em torno dos respectivos máximos (supostos únicos), tem-se

$$\ln h(\theta) = \ln h(m_0) - \frac{1}{2}(\theta - m_0)^t H_0 (\theta - m_0) + R_0$$

$$\ln f(x|\theta) = \ln f(x|\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^t H(\hat{\theta}_n)(\theta - \hat{\theta}_n) + R_n,$$

onde m_0 é a moda da distribuição *a priori*, $\hat{\theta}_n$ a estimativa de máxima verosimilhança de θ baseada nos dados x,

$$H_0 = -\left.\frac{\partial^2 \ln h(\theta)}{\partial\theta_i \partial\theta_j}\right|_{\theta=m_0}, \qquad H(\hat{\theta}_n) = -\left.\frac{\partial^2 \ln f(x|\theta)}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}_n}$$

e R_0, R_n, os restos dos respetivos desenvolvimentos em série. Admitindo certas condições de regularidade, que assegurem que estes restos sejam pequenos para grandes valores de n (ver, *e.g.*, Bernardo e Smith, 1994), tem-se

$$\begin{aligned} h(\theta|x) &\propto \exp\left\{-\frac{1}{2}(\theta - m_n)^t H_n (\theta - m_n)\right\}, \\ H_n &= H_0 + H(\hat{\theta}_n), \\ m_n &= H_n^{-1}(H_0 m_0 + H(\hat{\theta}_n)\hat{\theta}_n). \end{aligned} \tag{7.1}$$

Este desenvolvimento sugere que a distribuição *a posteriori* possa ser aproximada, para amostras de dimensão elevada e em condições gerais de regularidade, por uma distribuição Normal multivariada com valor médio m_n e matriz de covariâncias $\hat{\Sigma}_n = H_n^{-1}$ (simbolicamente escreve-se $N_k(m_n, \hat{\Sigma}_n)$).

À medida que a dimensão da amostra aumenta é de esperar que a precisão da distribuição *a priori*, representada por H_0, seja totalmente dominada pela precisão $H(\hat{\theta}_n)$ fornecida pelos dados, ou seja que $H_n \approx H(\hat{\theta}_n)$. Assim, também $m_n \approx \hat{\theta}_n$, podendo portanto usar-se como aproximação para a distribuição *a posteriori* a distribuição Normal multivariada centrada na estimativa de máxima verosimilhança e matriz de covariâncias $\hat{\Sigma} = \left[H(\hat{\theta}_n)\right]^{-1}$, o inverso da matriz de informação observada[31].

[31] Dado que a matriz de informação observada converge para a matriz de informação de Fisher, pode ainda usar-se como aproximação para a matriz de covariâncias o inverso da matriz de informação de Fisher.

Pode obter-se ainda outro tipo de aproximação se se considerar o desenvolvimento em série de $\ln h(\theta|x)$ em torno do seu ponto de máximo, isto é, em torno da moda da distribuição *a posteriori*. Suponha-se então que $\{h_n(\theta|x), n = 1, 2, \ldots\}$ é uma sucessão de distribuições *a posteriori* para θ. Seja $l_n(\theta|x) = \ln h_n(\theta|x)$ e m_n tal que

$$L'_n(m_n) = \partial l_n(\theta|x)/\partial\theta|_{\theta=m_n} = 0$$

e

$$\Sigma_n = (-L''_n(m_n))^{-1},$$

onde $[L''_n(m_n)]_{ij} = (\partial^2 l_n(\theta|x)/\partial\theta_i\partial\theta_j)|_{\theta=m_n}$.

Bernardo e Smith (1994, Cap. 5) provam que, sob certas condições sobre $h(\theta|x)$ e para valores grandes de n, a distribuição *a posteriori* pode ser aproximada por uma distribuição Normal multivariada com vetor média dado pela moda da distribuição *a posteriori* e matriz de covariâncias dada pelo negativo da inversa da matriz hessiana do logaritmo da densidade *a posteriori*, calculada na moda da distribuição.

Esta abordagem tem a vantagem de praticamente todas as formas de inferências sumárias *a posteriori* poderem ser calculadas segundo uma via gaussiana. Contudo, o problema maior advém da necessidade de averiguar, em cada aplicação, da adequabilidade da aproximação da distribuição *a posteriori* a uma distribuição Normal multivariada.

Exemplo 7.1 Suponha-se que X tem uma distribuição Binomial de parâmetros n (conhecido) e θ) e que se adotou para θ uma distribuição *a priori* conjugada Beta de parâmetros (a_0, b_0). Sabe-se que a distribuição *a posteriori* para θ é ainda uma Beta de parâmetros (a_n, b_n) onde $a_n = a_0 + x$ e $b_n = b_0 + n - x$. Assim

$$h_n(\theta|x) \propto \theta^{a_n-1}(1-\theta)^{b_n-1}, \quad 0 \le \theta \le 1.$$

Claro que se quiser fazer inferências sobre θ ou qualquer função de θ, por exemplo, sobre o logite, *i.e.*, $\rho = \ln\left(\frac{\theta}{1-\theta}\right)$, não há necessidade de recorrer a métodos computacionais sofisticados já que existem soluções exatas. Este exemplo é no entanto útil para exemplificar os métodos apresentados.

Tem-se

$$\ln h_n(\theta|x) \propto (a_n-1)\ln\theta+(b_n-1)\ln(1-\theta) \tag{7.2}$$

$$L_n'(\theta) = \frac{(a_n-1)}{\theta}-\frac{(b_n-1)}{(1-\theta)} \tag{7.3}$$

$$L_n''(\theta) = -\frac{(a_n-1)}{\theta^2}-\frac{(b_n-1)}{(1-\theta)^2} \tag{7.4}$$

e portanto

$$m_n=\frac{(a_n-1)}{a_n+b_n-2}, \qquad -\{L_n''(m_n)\}^{-1}=\frac{(a_n-1)(b_n-1)}{(a_n+b_n-2)^3}. \tag{7.5}$$

É fácil de verificar que as condições de regularidade são satisfeitas, e portanto, a distribuição *a posteriori* pode ser bem aproximada, para n grande, por uma distribuição Normal de valor médio m_n e variância dada por $\sigma_n^2 = -\{L_n''(m_n)\}^{-1}$ dados em (7.5).

No caso em que a distribuição *a priori* é a distribuição uniforme ($a_0 = b_0 = 1$) tem-se

$$m_n=\frac{x}{n}, \quad \sigma_n^2=\frac{x/n(1-x/n)}{n},$$

ou seja, a distribuição (*a posteriori*) de θ dado $X = x$ pode ser aproximada por uma distribuição $N(x/n,(x/n)(1-x/n)/n)$. Repare-se na dualidade deste resultado com o obtido, via teorema limite central, relativamente à distribuição assintótica de X/n dado θ. Com efeito, sabe-se que para $X \sim Bi(n,\theta)$, a distribuição de X/n dado θ é bem aproximada, para grandes valores de n, por uma distribuição $N(\theta,\theta(1-\theta)/n)$.

Se se quiser fazer inferências sobre ρ usando aproximações, pode proceder-se de vários modos. Comece-se por verificar que a distribuição *a posteriori* exata de ρ pode ser obtida por uma simples transformação. Com efeito, tem-se

$$h_n(\rho|x) \propto e^{a_n\rho}(1+e^{\rho})^{-(a_n+b_n)}, \quad \rho \in R \tag{7.6}$$

O valor médio *a posteriori* de ρ calculado a partir de (7.6) é $\psi(a_n)-\psi(b_n)$ onde a função $\psi(x)$ é a derivada do logaritmo da função $\Gamma(x)$[32].

[32]Ver Gradshteyn e Ryzhik, 1980, p. 943, para representações integrais e representações em série desta função.

Tem-se para esta distribuição

$$\ln h_n(\rho|x) \propto a_n\rho - (a_n+b_n)\ln(1+e^\rho) \tag{7.7}$$

$$L_n'(\rho) = a_n - (a_n+b_n)\frac{e^\rho}{(1+e^\rho)} \tag{7.8}$$

$$L_n''(\rho) = -(a_n+b_n)\frac{e^\rho}{(1+e^\rho)^2}, \tag{7.9}$$

obtendo-se

$$m_n = \ln\frac{a_n}{b_n} \qquad -\{L_n''(m_n)\}^{-1} = \frac{1}{a_n}+\frac{1}{b_n}. \tag{7.10}$$

Assim, a distribuição *a posteriori* para ρ pode ser aproximada por uma distribuição $N\left(\ln\frac{a_n}{b_n}, \frac{1}{a_n}+\frac{1}{b_n}\right)$. No caso de uma distribuição *a priori* vaga ($a_0 = b_0 = 0$) tem-se:

$$m_n = \ln\frac{x/n}{1-x/n}, \quad \sigma_n^2 = \frac{1}{n(x/n)(1-x/n)}.$$

Nesta situação, a distribuição (*a posteriori*) de ρ dado $X = x$ pode ser aproximada por uma distribuição $N\left(\ln\frac{x/n}{1-x/n}, \frac{1}{n(x/n)(1-x/n)}\right)$. Repare-se, novamente, na dualidade deste resultado com o obtido, via teorema limite central, para a distribuição assintótica de $\ln\frac{X/n}{1-X/n}$ dado θ. Com efeito, sabe-se que para $X \sim Bi(n,\theta)$, a distribuição de $\ln\frac{X/n}{1-X/n}$ dado θ é bem aproximada, para grandes valores de n, por uma distribuição $N\left(\ln\frac{\theta}{1-\theta}, \frac{1}{n\theta(1-\theta)}\right)$.

■

7.1.2 Método clássico de Laplace

Tierney e Kadane (1986) propuseram uma abordagem analítica para o cálculo de expressões da forma

$$E[g(\theta)|x] = \int g(\theta)h(\theta|x)d\theta, \tag{7.11}$$

usando o método de Laplace de aproximação de integrais. Este método consiste, sucintamente, no seguinte: Suponha-se que ψ é uma função regular de um parâmetro k-dimensional θ e que $-\psi$ tem um máximo em $\hat{\theta}$. O método de Laplace aproxima um integral da forma

$$I = \int f(\theta)\exp(-n\psi(\theta))d\theta, \tag{7.12}$$

através do desenvolvimento em série de ψ em torno de $\hat{\theta}$. Em geral, são suficientes desenvolvimentos até aos termos de segunda ordem. É o que se irá fazer a seguir.

- Considere-se primeiro o caso em que $k = 1$.

  Desenvolvendo $\psi(\theta)$ em torno de $\hat{\theta}$ até à segunda ordem e substituindo em $\exp(-n\psi(\theta))$ tem-se

  $$\exp(-n\psi(\theta)) \approx \exp\left(-n\psi(\hat{\theta}) - \frac{n(\theta-\hat{\theta})^2}{2}\psi''(\hat{\theta})\right),$$

  dado que $\psi'(\hat{\theta}) = 0$.

  Note-se que o termo da forma $\exp\left(-\frac{n\psi''(\hat{\theta})}{2}(\theta-\hat{\theta})^2\right)$, é proporcional à função densidade de probabilidade Normal de valor médio $\hat{\theta}$ e variância $(n\psi''(\hat{\theta}))^{-1}$. Consequentemente

  $$\int_{-\infty}^{+\infty} \exp\left(-\frac{n\psi''(\hat{\theta})}{2}(\theta-\hat{\theta})^2\right) = \left(2\pi(n\psi''(\hat{\theta}))^{-1}\right)^{\frac{1}{2}}.$$

  Assim o integral I em (7.12) pode ser aproximado por

  $$I \approx \hat{I}\{1 + O(n^{-1})\} \tag{7.13}$$

  onde

  $$\hat{I} = \sqrt{2\pi} n^{-\frac{1}{2}} \hat{\sigma} f(\hat{\theta}) \exp(-n\psi(\hat{\theta}))$$

  e $\hat{\sigma} = [\psi''(\hat{\theta})]^{-1/2}$.

- Para o caso k-dimensional um raciocínio idêntico conduz a $\hat{I}$ da forma:

  $$\hat{I} = (2\pi)^{\frac{k}{2}} n^{-\frac{k}{2}} \det(\hat{\Sigma})^{\frac{1}{2}} f(\hat{\theta}) \exp(-n\psi(\hat{\theta})),$$

  onde $\hat{\Sigma}^{-1} = \nabla^2\psi(\hat{\theta})$ é a matriz hessiana de ψ em $\hat{\theta}$.

Claro que, se nos desenvolvimentos de f e ψ se retiverem termos de ordem superior à segunda, se obtêm melhores aproximações, como por exemplo

$$\begin{aligned}\int f(\theta)e^{-n\psi(\theta)}d\theta &= \sqrt{(}2\pi)\sigma e^{-n\hat{\psi}}\left\{\hat{f} + \frac{1}{2n}\left[\sigma^2\hat{f}'' - \sigma^4\hat{f}'\hat{\psi}''' + \right.\right. \\ &+ \left.\left. \frac{5}{12}\hat{f}(\hat{\psi}''')^2\sigma^6 - \frac{1}{4}\hat{f}\hat{\psi}^{(4)}\sigma^4\right]\right\} + O(n^{-2}) \end{aligned} \tag{7.14}$$

onde $\hat{f}, \hat{\psi}$, etc., representam as respetivas funções calculadas em $\hat{\theta}$, que é, como já se referiu, o ponto onde $-\psi(\theta)$ atinge o máximo e $\sigma^2 = [\psi''(\hat{\theta})]^{-1}$.

Suponha-se então que se quer calcular o valor esperado *a posteriori* de uma função $g(\theta)$ do parâmetro. Por (7.11) vemos que $E[g(\theta)|x]$ se obtém como a razão de dois integrais, nomeadamente

$$E[g(\theta)|x] = \frac{\int g(\theta) f(x|\theta) h(\theta) d\theta}{\int f(x|\theta) h(\theta) d\theta}. \tag{7.15}$$

A ideia básica é aplicar separadamente o método de Laplace aos integrais do numerador e do denominador e considerar a razão das aproximações obtidas. Tierney e Kadane (1986) obtêm as seguintes duas aproximações para $E[g(\theta)|x]$, consoante o desenvolvimento utilizado.

- $E[g(\theta)|x] = g(\hat{\theta})[1 + O(n^{-1})]$

  Para se chegar a este resultado, aplique-se o método de Laplace fazendo $\exp(-n\psi(\theta)) = f(x|\theta)h(\theta)$, no numerador e denominador, $f(\theta) = g(\theta)$ no numerador e $f(\theta) = 1$ no denominador.

  Note-se que isto corresponde a aproximar $E[g(\theta)|x]$ pelo valor modal $g(\hat{\theta})$, onde $\hat{\theta}$ é a moda da distribuição *a posteriori*, já que $\hat{\theta}$ é o valor de θ que maximiza $-\psi(\theta)$.

- Supondo que $g(\theta)$ é positiva quase por toda a parte e, para simplificar, que θ é um parâmetro real, tem-se

  $$E[g(\theta)|x] = (\sigma^\star/\hat{\sigma}) \exp\{-n[\psi^\star(\theta^\star) - \psi(\hat{\theta})]\}(1 + O(n^{-2})).$$

  Para se chegar a esta aproximação considere-se

  $$E[g(\theta)|x] = \frac{\int \exp\{-n\psi^\star(\theta)\} d\theta}{\int \exp\{-n\psi(\theta)\} d\theta}, \tag{7.16}$$

  onde

  $$\begin{aligned} -n\psi(\theta) &= \ln h(\theta) + \ln f(x|\theta), \\ -n\psi^\star(\theta) &= \ln g(\theta) + \ln h(\theta) + \ln f(x|\theta). \end{aligned} \tag{7.17}$$

Defina-se $\hat{\theta}$, $\theta^\star$ e $\hat{\sigma}$, $\sigma^\star$ tal que

$$\begin{aligned} -\psi(\hat{\theta}) &= \sup_\theta\{-\psi(\theta)\}, \\ \hat{\sigma} &= [\psi''(\theta)]^{-1/2}|_{\theta=\hat{\theta}}, \\ -\psi^\star(\theta^\star) &= \sup_\theta\{-\psi^\star(\theta)\}, \\ \sigma^\star &= [\psi^{\star\prime\prime}(\theta)]^{-1/2}|_{\theta=\theta^\star}. \end{aligned} \tag{7.18}$$

Supondo que $\psi(\cdot)$, $\psi^\star(\cdot)$ são funções suficientemente regulares, as aproximações de Laplace para os integrais do numerador e denominador em (7.16) (fazendo em ambos os casos $f(\theta) = 1$) são, respetivamente,

$$\sqrt{2\pi}\sigma^\star n^{-1/2}\exp\{-n\psi^\star(\theta^\star)\},$$

e

$$\sqrt{2\pi}\hat{\sigma} n^{-1/2}\exp\{-n\psi(\hat{\theta})\}.$$

Daqui obtém-se a seguinte aproximação para $E[g(\theta)|x]$,

$$E[g(\theta)|x] \approx (\sigma^\star/\hat{\sigma})\exp\{-n[\psi^\star(\theta^\star) - \psi(\hat{\theta})]\}. \tag{7.19}$$

Os erros na aproximação dos dois integrais são da ordem n^{-1}. Contudo, os termos relevantes nos dois erros são idênticos e portanto, ao fazer a razão cancelam-se. Deste modo, a aproximação obtida tem agora um erro relativo da ordem n^{-2}.

Para se chegar à aproximação anterior impôs-se uma condição bastante restritiva, nomeadamente a condição de a função g ser positiva quase por toda a parte. Pode proceder-se de vários modos para se encontrar aproximações adequadas para a situação mais geral de uma função real g com contradomínio qualquer. Tierney *et al.* (1989) sugerem que se comece por calcular uma aproximação da função geradora de momentos de $g(\theta)$ ($E[\exp\{sg(\theta)\}]$) usando a aproximação de Laplace para funções positivas, e depois obter a aproximação do valor esperado de $g(\theta)$ como a derivada do logaritmo da função geradora em $s = 0$. O erro assim obtido é da ordem $O(n^{-2})$.

Outro processo, sugerido ainda por Tierney *et al.* (1989), é reescrever as funções integrandas em (7.15) de modo a que o valor esperado $E[g(\theta)|x]$ se possa escrever na forma

$$E[g(\theta)|x] = \frac{\int f_N(\theta)\exp\{-n\psi_N(\theta)\}d\theta}{\int f_D(\theta)\exp\{-n\psi_D(\theta)\}d\theta}, \tag{7.20}$$

para escolhas adequadas de f_N, f_D, ψ_N e ψ_D,[33] e usar como aproximação de Laplace para ambos os integrais a indicada em (7.14).

Exemplo 7.2 Voltando ao primeiro exemplo, se se quiser obter uma estimativa para $g(\theta) = \rho = \ln \frac{\theta}{1-\theta}$ pode usar-se o método de Laplace. Como $g(\theta)$ pode tomar valores negativos, ou se usa a primeira aproximação que corresponde a estimar o valor médio pelo valor modal, obtendo-se, portanto,

$$E(\rho|x) = \ln \frac{a_n - 1}{b_n - 1},$$

ou tem de se usar alguma das abordagens alternativas sugeridas por Tierney *et al.* (1989).

Usando, o desenvolvimento sugerido em (7.20) com $\psi_N = \psi_D = \psi$, sendo $-n\psi = \ln h(\theta) + \ln f(x|\theta)$; $f_N(\theta) = g(\theta)$; $f_D(\theta) = 1$, obtém-se

$$\begin{aligned} E(\rho|x) &= \ln \frac{a_n - 1}{b_n - 1} + \frac{1}{2n} \frac{(a_n - b_n)}{(a_n - 1)(b_n - 1)} \\ &- \frac{1}{n^2} \frac{(a_n - 1)(b_n - 1)}{(a_n + b_n - 2)} [(a_n - 1)^2 - (b_n - 1)^2]. \end{aligned} \tag{7.21}$$

Compare-se, especificando valores para a_n e b_n, estes resultados com o obtido no exemplo anterior e com o valor exato aí apresentado.

Tanner (1996) sugere obter uma aproximação ao valor médio de ρ através da transformação

$$\lambda = \frac{1}{2} \ln \frac{b_n \theta}{a_n (1 - \theta)}.$$

É fácil de ver que a distribuição de λ é a distribuição z de Fisher cuja densidade de probabilidade é

$$h(\lambda) \propto \frac{e^{2a_n \lambda}}{(2b_n + 2a_n e^{2\lambda})^{(a_n + b_n)}},$$

sendo o seu valor médio aproximado

$$\frac{1}{2} \ln \left[\frac{1 - (2a_n)^{-1}}{1 - (2b_n)^{-1}} \right].$$

Assim obtém-se para valor médio aproximado para ρ

$$\ln \frac{a_n - 0.5}{b_n - 0.5},$$

[33] Pode usar-se, por exemplo, $\psi_N = \psi_D; f_N(\theta) = g(\theta); f_D(\theta) = 1$. Ver Tierney *et al.* (1989) ou Robert (1994) para mais pormenores.

que é mais preciso que os anteriores. ■

As ideias são facilmente estendidas para o caso multiparamétrico tendo aplicação direta no cálculo de densidades *a posteriori* marginais, momentos *a posteriori* e densidades preditivas.

Suponha-se que $\theta \in R^k$ e que se pretende determinar a distribuição marginal *a posteriori* de θ_1. Decomponha-se $\theta = (\theta_1, \theta_{-(1)})$ onde $\theta_{-(1)} = (\theta_2, ..., \theta_k)$. A distribuição marginal *a posteriori* de θ_1 pode escrever-se como a razão de dois integrais

$$h_1(\theta_1|x) = \int_{\theta_{-(1)}} h(\theta_1, \theta_{-(1)}|x) d\theta_{-(1)} \tag{7.22}$$

$$= \frac{\int_{\theta_{-(1)}} h(\theta_1, \theta_{-(1)}) f(x|\theta_1, \theta_{-(1)}) d\theta_{-(1)}}{\int_\theta h(\theta) f(x|\theta) d\theta}. \tag{7.23}$$

Aplicando o método de Laplace ao numerador e denominador de (7.22) obtém-se a aproximação

$$h_1(\theta_1|x) \approx \left(\frac{\det(\hat{\Sigma}^\star(\theta_1))}{2\pi n \det(\hat{\Sigma})} \right)^{1/2} \frac{h(\theta_1, \hat{\theta}_{-(1)}) f(x|\theta_1, \hat{\theta}_{-(1)})}{h(\hat{\theta}) f(x|\hat{\theta})} \tag{7.24}$$

onde $\hat{\theta}$ maximiza $h(\theta)f(x|\theta)$, sendo $\hat{\Sigma}$ o negativo do inverso da correspondente matriz hessiana calculada em $\hat{\theta}$, $\hat{\theta}_{-(1)}$ maximiza $h(\theta_1, \theta_{-(1)}) f(x|\theta_1, \theta_{-(1)})$ para θ_1 fixo, sendo $\hat{\Sigma}^\star(\theta_1)$ o negativo da correspondente matriz hessiana calculada em $\hat{\theta}_{-(1)}$.

Veremos na secção 7.3 como este resultado irá ser útil na metodologia INLA desenvolvida por Rue e seus colaboradores.

Pode ainda mostrar-se, usando argumentos semelhantes, que (7.19) é ainda válida quando $\theta \in R^k$ com

$$\hat{\sigma} = |\nabla^2 \psi(\hat{\theta})|^{-1/2} \qquad \text{e} \qquad \sigma^\star = |\nabla^2 \psi^\star(\theta^\star)|^{-1/2},$$

onde

$$[\nabla^2 \psi(\theta)]_{ij} = \frac{\partial^2 \psi(\theta)}{\partial \theta_i \partial \theta_j} \qquad \text{e} \qquad [\nabla^2 \psi^\star(\theta)]_{ij} = \frac{\partial^2 \psi^\star(\theta)}{\partial \theta_i \partial \theta_j}.$$

Para mais informações sobre este método, veja-se referências citadas no livro Paulino *et al.* (2003). Embora esta metodologia represente uma técnica potente, tem alguns inconvenientes. No caso multiparamétrico a sua aplicação pode ser difícil e pouco prática, principalmente se as funções integrandas são

multimodais e/ou as derivadas locais são difíceis de obter. Em muitas situações, mesmo quando a dimensão do parâmetro não é elevada, é conveniente fazer reparametrizações de modo a obter melhores aproximações.

7.2 Modelos gaussianos latentes (LGM)

A classe dos denominados modelos gaussianos latentes pode ser representada por uma estrutura hierárquica em três níveis. O primeiro nível corresponde à verosimilhança condicional

$$X|\theta,\psi \sim f(x|\theta,\psi) = \prod_{i=1}^{n} f(x_i|\theta,\psi) \tag{7.25}$$

O segundo nível corresponde a admitir que o vetor de parâmetros θ é governado por um campo aleatório gaussiano com estrutura markoviana (*Gaussian Markov Random Field*, GMRF), com respeito a um grafo não direcionado $\mathcal{G} = (\mathcal{V} = \{1,...,n\}, \mathcal{E})$ (ver Rue e Held (2005)), isto é

$$\theta|\psi \sim N(0, \Sigma(\psi)) \tag{7.26}$$

$$\theta_l \perp\!\!\!\perp \theta_m|\theta_{-(lm)}, \qquad \forall\{l,m\} \notin \mathcal{E} \tag{7.27}$$

onde $\theta_{-(lm)}$ corresponde ao vetor θ excluindo as componentes θ_l e θ_m, significando pois que θ_l e θ_m que não tenham arestas comuns são condicionalmente independentes.

Como a matriz de covariância de θ depende de hiperparâmetros ψ, o terceiro nível corresponde à especificação da distribuição *a priori* para ψ.

Em muitas situações tem-se que o vetor de hiperparâmetros ψ pode ser particionado em $\psi = (\psi_1, \psi_2)$ de tal modo que o LGM pode ser reexpresso como

$$\begin{aligned}
\psi &\sim h(\psi) && \text{(distribuição } a \textit{ priori} \text{ dos hiperparâmetros)} \\
\theta|\psi &\sim N(0, \Sigma(\psi_1)) && \text{(GMRF como distribuição } a \textit{ priori} \text{ para } \theta) \\
x|\theta,\psi &\sim f(x|\theta,\psi) = \prod_{i=1}^{n} f(x_i|\theta_i,\psi_2) && \text{(modelo amostral para os dados } x)
\end{aligned}$$

Neste caso ψ_1 é o vetor de hiperparâmetros e ψ_2 é um vetor de parâmetros perturbadores. Na formulação destes modelos o vetor θ pode ter uma dimensão muito elevada, contrariamente ao vetor ψ que deve ter uma dimensão geralmente pequena (1 a 5).

Exemplo 7.3 Num estudo longitudinal, n doentes com a mesma doença foram submetidos a dois tratamentos diferentes. A avaliação clínica após a aplicação do tratamento foi feita em três tempos distintos. O objetivo do estudo é saber se a evolução da doença é diferente nos dois grupos, se depende da idade do doente e da duração da doença antes do tratamento e de uma covariável medida também nos três tempos em que é feita a avaliação clínica da evolução da doença.

O tratamento bayesiano deste modelo de regressão com medições repetidas enquadra-se facilmente dentro da formulação dos LGM. Com efeito seja X_{jk} a variável aleatória que representa o resultado da avaliação clínica de um doente j (com $j = 1, ..., n$) no instante k (com $k = 1, 2, 3$). Seja ainda $z_j = (z_{1,j}, z_{2,j}, z_{3,j})$ o vetor de covariáveis tratamento, idade e duração da doença, respetivamente, do paciente j e z_{jk} a covariável que varia no tempo.

Admita-se que se considera o seguinte modelo distribucional para X_{jk}:

- $X_{jk} \sim N(\mu_{jk}, \sigma^2)$ com independência condicional (aos seus parâmetros)
- $\mu_{jk} = \beta_0 + \beta_1 z_{1,j} + \beta_2 z_{2,j} + \beta_3 z_{3,j} + \beta_4 z_{jk} + a_j + b_{jk}$
- $a_j \underset{iid}{\sim} N(0, \sigma_a^2)$, $b_{jk} \underset{iid}{\sim} N(0, \sigma_b^2)$ de modo que $a = (a_j, j = 1, ..., n)$ e $b = (b_{jk}, j = 1, ..., n, k = 1, 2, 3)$ representam efeitos aleatórios a nível dos indivíduos e a nível das observações dentro dos indivíduos. Estes efeitos aleatórios são introduzidos para acomodar a dependência existente devido à natureza longitudinal dos dados.
- $\beta = (\beta_1, \beta_2, \beta_3, \beta_4), \quad \beta_0, \beta_i, i = 1, \ldots, 4 \underset{iid}{\sim} N(0, \sigma_\beta^2)$
- $\tau = (\sigma^2)^{-1} \sim Ga(c, d), \tau_a = (\sigma_a^2)^{-1} \sim Ga(c_a, d_a), \tau_b = (\sigma_b^2)^{-1} \sim Ga(c_b, d_b)$, $\tau_\beta = (\sigma_\beta^2)^{-1} \sim Ga(c_\beta, d_\beta)$

Este modelo hierárquico apresenta-se em três níveis:

1. $X|z, \theta, \psi \sim f(x|z, \theta, \psi) = \prod_{j,k} N(x_{jk}|z_j, \mu_{jk}, \psi_2)$
2. $\theta = (\beta_0, \beta, a, b)$ com $\sim N(0, \Sigma(\psi_1))$, ou seja $\theta|\psi \sim GMRF(\psi_1)$
3. $\psi = (\psi_1, \psi_2)$ sendo $\psi_1 = (\tau_\beta, \tau_a, \tau_b)$ no caso em que se admite os parâmetros do modelo como independentes *a priori* e $\psi_2 = \tau$.

■

Note-se que a independência *a priori*, acima admitida, no que diz respeito aos parâmetros relativos aos efeitos fixos corresponde a uma simplificação do problema que não é necessária. Uma das propriedades interessantes dos GMRF (ver Rue e Held, 2005) é que a matriz de precisão $Q = \Sigma^{-1}$ é uma matriz esparsa. Com efeito prova-se que para um GMRF

$$\theta_l \perp\!\!\!\perp \theta_m | \theta_{-(lm)} \Leftrightarrow Q_{lm} = 0,$$

onde Q_{lm} representa a componente lm da matriz de precisão Q. Esta propriedade traz como vantagem a redução de esforço computacional devido à possibilidade da utilização de métodos numéricos específicos para lidar com matrizes esparsas.

Um modo alternativo de formular os modelos LGM, fundamental na abordagem INLA, é através do seu enquadramento como subclasse dos modelos de regressão aditiva estruturada (Fahrmeir e Tutz, 2001).

Nesta formulação a variável dependente (X_i) pertence à família exponencial onde a média μ_i está ligada a um preditor η_i com estrutura aditiva através de uma função de ligação $g(\mu_i) = \eta_i$, podendo a verosimilhança do modelo ser ainda controlada por um hiperparâmetro ψ_2. A forma geral do preditor é

$$\eta_i = \beta_0 + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} w_{ki} f^{(k)}(u_{ki}) + \epsilon_i, \tag{7.28}$$

onde β_0 é a ordenada na origem, $\beta = (\beta_1, ..., \beta_{n_\beta})$ é o vetor que representa os efeitos lineares das covariáveis z. As funções $(f^{(1)}, ..., f^{(n_f)})$ de covariáveis u podem representar efeitos não lineares de covariáveis contínuas, efeitos sazonais, efeitos aleatórios estruturados de natureza diversa. Estas funções podem ter pesos associados $(\{w_{ki}\})$ conhecidos para cada observação. Efeitos aleatórios não estruturados são contemplados através do termo ϵ_i.

Um modelo gaussiano latente é obtido atribuindo a $\theta = \{\beta_0, \{\beta_j\}, \{f^{(k)}\}, \{\eta_i\}\}$ uma distribuição *a priori* gaussiana com matriz de precisão $Q(\psi_1)$ (mais especificamente um GMRF). Esta parametrização de θ, que inclui η_i na sua composição, é útil dado que permite associar a cada observação uma componente do campo aleatório.

A definição do modelo latente fica completa com a atribuição de uma distribuição a priori aos hiperparâmetros do modelo (ψ_1, ψ_2).

7.3 Abordagem via aproximações de Laplace encaixadas e integradas (INLA)

Usando a formulação dos modelos gaussianos latentes, a distribuição *a posteriori* de interesse

$$\begin{aligned} h(\theta,\psi|x) &\propto h(\theta|\psi)h(\psi)\prod_i f(x_i|\theta_i,\psi) \\ &\propto h(\psi)|Q(\psi)|^{n/2}\exp\left(-\frac{1}{2}\theta^T Q(\psi)\theta + \sum_i \log(f(x_i|\theta_i,\psi))\right) \end{aligned}$$

O objetivo principal com a abordagem INLA é a obtenção de uma aproximação analítica para distribuições marginais *a posteriori* dos parâmetros latentes e dos seus hiperparâmetros no modelo gaussiano.

Tais marginais podem ser escritas como:

$$\begin{aligned} h(\theta_i|x) &= \int h(\theta_i,\psi|x)d\psi = \int h(\psi|x)h(\theta_i|\psi,x)d\psi \\ h(\psi_k|x) &= \int h(\psi|x)d\psi_{-(k)}, \end{aligned}$$

onde $\psi_{-(k)}$ denota o vetor ψ sem a componente ψ_k. Para o cálculo destas distribuições precisa-se de estimativas prévias de $h(\psi|x)$ e $h(\theta_i|\psi,x)$ para então obter

$$\tilde{h}(\theta_i|x) = \int \tilde{h}(\psi|x)\tilde{h}(\theta_i|\psi,x)d\psi \tag{7.29}$$

$$\tilde{h}(\psi_k|x) = \int \tilde{h}(\psi|x)d\psi_{-(k)}, \tag{7.30}$$

onde $\tilde{h}(.|.)$ representa uma aproximação à densidade respetiva. Assim, a obtenção de aproximações para as distribuições marginais *a posteriori* (7.29)-(7.30) passa por três etapas: aproximação para $h(\theta_i|\psi,x)$, aproximação para $h(\psi|x)$ e integração numérica. A designação INLA advém precisamente do conjunto destas etapas. Para além de uma etapa de integração, a aproximação de Laplace $h(\theta|\psi,x)$ é usada para obter a aproximação de Laplace para $h(\psi|x)$, a qual por sua vez é usada para obter uma aproximação para as distribuições marginais pretendidas.

É fácil de verificar que

$$h(\psi|x) = \frac{h(\theta,\psi|x)}{h(\theta|\psi,x)} \propto \frac{h(\psi)h(\theta|\psi)f(x|\theta,\psi)}{h(\theta|\psi,x)}.$$

Se $\widetilde{h}(\theta|\psi,x)$ fôr uma aproximação gaussiana para $h(\theta|\psi,x)$ e $\widehat{\theta}(\psi)$ a sua moda, uma aproximação para $h(\psi|x)$ pode ser obtida seguindo o método de Laplace proposto por Tierney e Kadane (1986) para obter aproximações à distribuição *a posteriori* marginal e já referida em (7.24),

$$\widetilde{h}(\psi|x) \propto \left.\frac{h(\psi)h(\theta|\psi)f(x|\theta,\psi)}{\widetilde{h}(\theta|\psi,x)}\right|_{\theta=\widehat{\theta}(\psi)}.$$

Dado que

$$h(\theta|\psi,x) \propto \exp\left\{-\frac{1}{2}\theta^T Q\theta - \sum_i \log(f(x_i|\theta_i,\psi))\right\},$$

pode obter-se uma aproximação gaussiana para $h(\theta|\psi,x)$ através de um processo iterativo, considerando o desenvolvimento em série de Taylor até à segunda ordem de $\log(f(x_i|\theta_i,\psi)) = g_i(\theta_i|\psi)$ em torno da i-ésima componente $\mu_i^{(0)}(\psi)$ de um valor inicial para o vetor média $\mu^{(0)}(\psi)$. Detalhes podem ser vistos em Rue *et al.* (2009).

Relativamente à obtenção de uma aproximação para $h(\theta_i|\psi,x)$ há várias abordagens possíveis a considerar:

1. Usar diretamente uma aproximação à Normal para $h(\theta_i|\psi,x)$ e uma decomposição de Cholesky para a matriz de precisão $Q(\psi_1)$, isto é, $Q(\psi_1) = L(\psi_1)L^T(\psi_1)$, onde $L(\psi_1)$ é uma matriz triangular inferior, para a obtenção das variâncias marginais. Deste modo só há o trabalho adicional de calcular as variâncias marginais. No entanto, esta aproximação à Normal para $h(\theta_i|\psi,x)$ não é geralmente muito boa.

2. Seja θ_{-i} o vetor θ excluindo a componente θ_i; tem-se

$$h(\theta_i|\psi,x) = \frac{h(\theta_i,\theta_{-i}|\psi,x)}{h(\theta_{-i}|\theta_i,\psi,x)} \propto \frac{h(\psi)h(\theta|\psi)f(x|\theta,\psi)}{h(\theta_{-i}|\theta_i,\psi,x)}.$$

   Usando uma aproximação gaussiana para $h(\theta_{-i}|\theta_i,\psi,x)$, uma estimativa para $h(\theta_i|\psi,x)$ é então

$$\tilde{h}(\theta_i|\psi,x) \propto \left.\frac{h(\psi)h(\theta|\psi)f(x|\theta,\psi)}{\tilde{h}(\theta_{-i}|\theta_i,\psi,x)}\right|_{\theta_{-i}=\widehat{\theta_{-i}}(\theta_i,\psi)}, \tag{7.31}$$

   onde $\widehat{\theta_{-i}}(\theta_i,\psi)$ é a moda de $\tilde{h}(\theta_{-i}|\theta_i,\psi,x)$. Esta abordagem dá melhores resultados que a anterior no que diz respeito à aproximação gaussiana,

mas apresenta o inconveniente desta aproximação ter de ser recalculada para todo o θ_i e ψ pois a matrix de precisão depende de θ_i e ψ.

3. Para ultrapassar esta dificuldade, Rue *et al.* (2009) sugeriram várias modificações dando origem a procedimentos alternativos que designaram por abordagem de Laplace completa e abordagem de Laplace simplificada.

   Na abordagem completa, evita-se o processo de otimização usando em vez da moda o valor médio condicional $E(\theta_{-i}|\theta_i)$ deduzido da aproximação gaussiana $\tilde{h}(\theta|\psi, x)$. Para além disso, na obtenção da distribuição marginal para θ_i só são considerados os θ_j que lhe são "próximos", considerando que apenas esses devem ter impacto na distribuição *a posteriori* marginal de θ_i.

   Considera-se que um θ_j está próximo de θ_i se $|a_{ij}(\psi)| > 0.001$, onde a_{ij} é obtido de

   $$\frac{E(\theta_j|\theta_i) - \mu_j(\psi)}{\sigma_j(\psi)} = a_{ij}(\psi)\frac{\theta_i - \mu_i(\psi)}{\sigma_i(\psi)},$$

   onde o valor médio condicional, $\mu_i, \sigma_i, \mu_j, \sigma_j$, são obtidos da aproximação gaussiana $\tilde{h}(\theta|\psi, x)$.

   A versão simplificada baseia-se desenvolvimento em série de Taylor até à terceira ordem do logaritmo tanto do numerador como do denominador de (7.31), usando novamente para θ_{-i} a esperança condicional em vez da moda. No numerador, os termos de terceira ordem permitem corrigir a aproximação no que diz respeito à assimetria. Detalhes destes procedimentos podem ser encontrados em Rue *et al.* (2009).

Como é natural, os detalhes computacionais no desenvolvimento destes métodos não são triviais. Uma descrição tanto quanto possível exaustiva de parte desses detalhes pode ser encontrada em Rue *et al.* (2009) e Blangiardo *et al.* (2013). A implementação eficaz destes procedimentos é conseguida pelo pacote computacional de livre acesso R-INLA (consultar www.r-inla.org).

Capítulo 8

Software

Métodos computacionais, referidos nos capítulos anteriores, para a análise de dados com recurso à metodologia bayesiana permitiram a aplicação generalizada desta metodologia a problemas oriundos de uma grande diversidade de áreas científicas. Para este desenvolvimento muito tem contribuído a generosidade de muitos investigadores que têm posto livremente à disposição da comunidade científica software para a implementação desta metodologia. O software `R` possui uma variedade de pacotes, ou por vezes só funções, que podem ser utilizados para fazer inferência bayesiana. Aconselha-se o leitor interessado a consultar a página http://cran.r-project.org/web/views/Bayesian.html. Aí pode-se encontrar, por exemplo, o pacote `DPpackage` que contém funções para fazer inferência bayesiana não paramétrica, o pacote `bayesSurv` específico para fazer inferência bayesiana em modelos de sobrevivência, etc. Do software que implementa métodos baseados em simulação estocástica, escolheu-se inserir neste capítulo quatro tipos de software generalista de acesso livre e que, embora independentes do `R`, podem ser utilizados através de ligação ao `R`. São eles o `OpenBUGS` (Thomas et *al.*, 2006), `JAGS` (Plummer, 2003), `Stan` (Stan Development Team, 2014 a) e `BayesX` (Adler *et al.*, 2013 e Belitz *et al.*, 2013). Faz-se uma apresentação sumária das principais caraterísticas do software e exemplifica-se a sua utilização com recurso ao mesmo exemplo. Mostra-se ainda como a monitorização da convergência das cadeias pode ser feita com recurso ao software CODA e BOA, ambos pacotes do `R`. Por fim mostra-se como esse exemplo pode ser tratado com recurso ao `R-INLA` (consultar www.r-

inla.org).

Atualmente há vários livros que ilustram a utilização destes software. Refere-se aqui os livros de Ntzoufras (2009), Kruschke (2011, 2014), Korner-Nievergelt *et al.* (2015) e Blangiardo e Cameletti (2015).

8.1 Exemplo de aplicação

Por comodidade relembra-se aqui o exemplo a ser tratado com os diferentes métodos.

Exemplo 8.1 Num estudo longitudinal, n doentes com a mesma doença, foram submetidos a dois tratamentos diferentes. A avaliação clínica após a aplicação do tratamento foi feita em três tempos distintos. O objetivo do estudo é saber se a evolução da doença é diferente nos dois grupos e se depende da idade do doente, da duração da doença antes do tratamento e de uma covariável medida também nos três tempos em que é feita a avaliação clínica da evolução da doença.

O tratamento bayesiano do modelo de regressão com medições repetidas a seguir considerado enquadra-se facilmente dentro da formulação dos LGM. Com efeito, seja X_{jk} a variável aleatória que representa o resultado da avaliação clínica de um doente j (com $j = 1, ..., n$) no instante k (com $k = 1, 2, 3$). Seja ainda $z_j = (z_{1,j}, z_{2,j}, z_{3,j})$ o vetor de covariáveis tratamento, idade e duração da doença, respetivamente, do paciente j e z_{jk} a covariável que varia no tempo.

Admita-se o seguinte modelo distribucional para as variáveis X_{jk}:

1. $X_{jk} \sim N(\mu_{jk}, \sigma^2)$ com independência condicional (aos seus parâmetros);
2. $\mu_{jk} = \beta_0 + \beta_1 z_{1,j} + \beta_2 z_{2,j} + \beta_3 z_{3,j} + \beta_4 z_{jk} + a_j + b_{jk}$.
3. $a_j \underset{iid}{\sim} N(0, \sigma_a^2)$, $b_{jk} \underset{iid}{\sim} N(0, \sigma_b^2)$ de modo que $a = (a_j, j = 1, ..., n)$ e $b = (b_{jk}, j = 1, ..., n, k = 1, 2, 3)$ representam efeitos aleatórios ao nível dos indivíduos e das observações dentro dos indivíduos, respetivamente. Estes efeitos aleatórios são introduzidos para acomodar a dependência existente devido à natureza longitudinal dos dados.
4. $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$, $\quad \beta_0, \beta_i, i = 1, \ldots, 4 \underset{iid}{\sim} N(0, \sigma_\beta^2)$.

5. $\tau = (\sigma^2)^{-1} \sim Ga(c, d), \tau_a = (\sigma_a^2)^{-1} \sim Ga(c_a, d_a), \tau_b = (\sigma_b^2)^{-1} \sim Ga(c_b, d_b)$, $\tau_\beta = (\sigma_\beta^2)^{-1} \sim Ga(c_\beta, d_\beta)$.

■

8.2 O projeto BUGS: WinBUGS e OpenBUGS

O projeto BUGS (*Bayesian inference using Gibbs sampling*) iniciou-se em 1989 com a contratação de Andrew Thomas para a *Medical Research Council Biostatistics Unit* em Cambridge. O projeto surgiu na sequência de trabalhos que vinham a ser desenvolvidos por Spiegelhalter (1986) e Lauritzen e Spiegelhalter (1988) sobre probabilidades em estruturas gráficas e sua aplicação em inteligência artificial, e o reconhecimento da importância dessas estruturas na formulação de modelos bayesianos. É interessante notar que, simultaneamente, mas independentemente, o trabalho seminal de Gelfand e Smith (1990) estava a ser desenvolvido em Nottingham, embora através de uma perspetiva bastante diferente.

Um protótipo do software BUGS foi apresentado pela primeira vez em público no *IV Valencia Meeting* em 1991. De acordo com Lunn *et al* (2009), o verdadeiro ímpeto para o seu desenvolvimento deu-se em 1993 após o *INSERM workshop on MCMC methods*, a que se seguiu um *workshop* em Cambridge sobre utilização do software BUGS, tendo dado origem à publicação do livro de Gilks, Richardson e Spiegelhalter sobre MCMC in Practice em 1996. Nas primeiras versões do software, apenas era utilizado o amostrador de Gibbs e para modelos em que as distribuições condicionais completas eram log-côncavas. A grande evolução do software deu-se a partir de 1996, quando o projeto se moveu para o Imperial College em Londres. Foi criada a versão WinBUGS permitindo fazer diagnósticos e inferências de um modo interativo. A introdução da utilização de métodos de simulação de Metropolis-Hastings permitiu o abandono da exigência da log-concavidade dos modelos a serem analisados.

A análise de modelos mais complexos foi também conseguida com a introdução do amostrador em fatias (*slice sampler*) e o WinBUGS *Jump Interface* para implementar o algoritmo de MCMC com saltos reversíveis. Foi desenvolvido o GeoBUGS para análise de dados espaciais, o PKBugs para a análise de modelos farmacocinéticos e o WBDiff para lidar com sistemas de equações diferenciais ordinárias. Em 2004 Andrew Thomas começou a trabalhar numa

versão *open-source* do software BUGS na Universidade de Helsinkia, dando origem ao projeto OpenBUGS, o qual se espera que seja o futuro do projeto BUGS já que em 2007 foi lançada a última versão do WinBUGS, a versão 1.4.3, com uma chave universal. Uma história do desenvolvimento do projeto, incluindo a sua evolução técnica, pode ser lida em Lunn *et al.* (2009). Cabe, com certeza, à disponibilização do software BUGS, parte da responsabilidade da enorme disseminação das ideias bayesianas durante o fim do século passado e da utilização generalizada da metodologia bayesiana durante a primeira década deste século, tal como foi preconizado por Lindley. O software BUGS gozou durante muitos anos da ausência de competidores, apesar dos reconhecidos defeitos deste software, muito particularmente o tempo computacional exigido.

A linguagem BUGS é simples e atrativa dado que corresponde a uma definição textual do modelo bayesiano. As relações estocásticas são representadas pelo símbolo ~ e as relações lógicas e/ou determinísticas pelo símbolo < −. Inclui estruturas típicas de linguagens de programação como o ciclo `for`. Como linguagem declarativa que é, a ordem das instruções é irrelevante. Isto tem o inconveniente de não permitir, por exemplo, o uso de instruções como *if-then-else*, o que corresponde a uma grande limitação da linguagem. A função *step*() pode ser usada para ultrapassar esta limitação, mas a escrita do programa acaba por se tornar pouco clara.

Dada a descontinuidade do WinBUGS, atualmente é aconselhável utilizar-se o OpenBUGS. O pacote R2OpenBUGS, uma adaptação de Neal Thomas do pacote R2WinBUGS (Sturtz *et al.*, 2005), serve de interface entre o OpenBUGS e o R. Para analisar o modelo bayesiano no R através do R2OpenBUGS é necessário, em primeiro lugar, escrever o código do programa do modelo estatístico na linguagem BUGS. Aliás, é útil fazê-lo no próprio OpenBUGS pois há a oportunidade de verificar a sintaxe do programa e corrigir erros eventuais.

Há vários manuais relativos ao BUGS, os quais podem ser descarregados do sítio http://www.openbugs.net/w/Manuals. O manual básico é o OpenBUGS User Manual, o qual deve ser consultado para perceber como se definem os modelos e toda a funcionalidade do BUGS. Uma lista das distribuições de amostragem e distribuições *a priori* aceites, e como as especificar, encontra-se no apêndice I desse manual. No caso de a distribuição de amostragem não se encontrar nessa lista, há a possibilidade de o utilizador construir a sua própria verosimilhança. Na secção desse manual *Advanced Use of the BUGS Language* explica-se como fazê-lo. O mesmo tipo de procedimento aplica-se à

especificação de uma distribuição *a priori* para um parâmetro que não faça parte da lista das distribuições *a priori* estabelecidas no dito apêndice.

Ilustra-se em seguida a utilização do exemplo referido no introito deste capítulo.

8.2.1 Exemplo de aplicação: recurso ao R2OpenBUGS

Depois de abrir a sessão no R deve instalar-se o software usando a instrução

```
install.packages("R2OpenBUGS", dependencies=TRUE,
repos="http://cran.us.r-project.org")
```

Para simplificar notação vai-se sempre admitir que todos os ficheiros necessários para a execução dos programas estão guardados no diretório em que foi aberta a sessão de R.

Os passos necessários para executar o modelo no R2OpenBUGS são os seguintes:

1. Escrever o código do modelo e guardá-lo num ficheiro com a extensão .txt. Neste caso foi guardado com o nome Cexemplo1BUGS.txt.

```
model{
for(i in 1:147){
X[i]~dnorm(mu[i],tau)
mu[i]<-beta0+beta[1]*z1[i]+beta[2]*z2[i]
+beta[3]*z3[i]+beta[4]*z[i]+a[ID[i]]+b[i]
b[i]~dnorm(0,tau_b)
}
for(j in 1:49){
a[j]~dnorm(0,tau_a)
}
for(k in 1:4){
beta[k]~dnorm(0,0.0001)
}
beta0~dnorm(0,0.0001)
tau~dgamma(0.05,0.05)
tau_a~dgamma(0.05,0.05)
tau_b~dgamma(0.05,0.05)
sigma<-1/sqrt(tau)
```

```
sigma_a<-1/sqrt(tau_a)
sigma_b<-1/sqrt(tau_b)
}
```

O primeiro ciclo de `for` corresponde à formulação do modelo probabilístico indicado no ponto (1) da secção 8.1. No `BUGS` o segundo parâmetro (*tau*) da distribuição é a precisão (inverso da variância). Neste ciclo está também incluída a definição do preditor linear *mu* e a formulação do modelo para os efeitos aleatórios b que aparecem nos pontos (2) e (3) da mesma secção. O número de doentes é $n = 49$. Dado que foram observados em três tempos distintos, o número total de observações é 147.

O segundo ciclo de `for` corresponde à formulação do modelo para os efeitos aleatórios a. Estes efeitos são específicos dos indivíduos que estão identificados por um índice com o símbolo ID.

O terceiro ciclo de `for` corresponde à formulação da distribuição *a priori* para os parâmetros dos efeitos fixos. Seguidamente está formulada a distribuição *a priori* dos restantes parâmetros do modelo. Todas as distribuições *a priori* formuladas refletem a natureza vaga da informação *a priori* existente. Para poder monitorizar os desvios padrões eles são definidos, por fim, em função das precisões respetivas. De notar que a ordem destas instruções é perfeitamente arbitrária.

Para se verificar o ficheiro que define o modelo a ser usado pelo `OpenBUGS` pode escrever-se a instrução

```
file.show("Cexemplo1BUGS.txt")
```

2. Fazer a leitura dos dados.

É aconselhável que as covariáveis que aparecem na definição do preditor linear sejam centradas na correspondente média para efeitos de aceleração do processo de convergência das cadeias. Se o ficheiro dos dados não estiver construído desse modo, esse procedimento pode ser depois feito no `R`.

```
> Cexemplo1<-read.table("Cexemplo1.txt",header=T)
> dim(Cexemplo1)
[1] 147  11
> head(round(Cexemplo1,3))

   X  gender z1    z2      year   z3   ID     z     all
```

```
1 14      1  0  -0.89796    1   2.5979  1  19.1329   1
2 10      2  0  -0.89796    1  -0.4020  2 -13.0671   2
3  8      2  0   0.10204    1  -0.9020  3  -7.3671   3
4 10      2  0  -3.89796    1  -1.2020  4  51.2329   4
5 10      2  0  -7.89796    1  -1.7020  5  18.2329   5
6 20      2  0  -3.89796    1   0.5979  6  -0.8671   6
```

A covariável `z` foi medida nos três tempos de avaliação. A variável `ID`, varia de 1 a 49, sendo um identificador do indivíduo. A variável `all` varia de 1 a 147 e aparece no ficheiro por conveniência, como se verá mais tarde. A variável `year` varia de 1 a 3 e identifica o período de avaliação. Também aparece no ficheiro por conveniência. Neste ficheiro as covariáveis contínuas, `z2, z3` e `z` já se encontram centradas. A variável `z1` toma os valores 0 e 1, consoante o tratamento atribuído ao doente.

3. Definir os vetores da matriz dos dados a usar no modelo. Os dados têm de ser fornecidos na forma de vetor, matriz ou lista.

```
#Criação de objetos separados para cada variável
X<-Cexemplo1$X
ID<-Cexemplo1$ID
z3<-Cexemplo1$z3
z1<-Cexemplo1$z1
z2<-Cexemplo1$z2
z<-Cexemplo1$z
#Criação de uma lista com os dados que serão fornecidos ao OpenBUGS
Cexemplo1.data<-list("X","ID","z1","z2","z3","z")
```

4. Definir os parâmetros que se pretende monitorizar.

```
Cexemplo1.params <-
c("beta0","beta","tau","tau_a","tau_b","sigma_a","sigma_b","sigma")
```

Se se quisesse monitorizar o "mu", ou quaisquer outros parâmetros que tivessem sido definidos no corpo do modelo, por exemplo, "a", ou "b", eles também deveriam aparecer na lista acima.

5. Definir os valores iniciais dos parâmetros e hiperparâmetros do modelo. Neste caso há que definir valores iniciais para "beta0", "beta", "tau", "tau_a", "tau_b", "a" e "b". Eles têm de ser definidos através de uma lista, devendo aparecer uma lista de valores iniciais para cada cadeia.

```
Inits<-list(tau=1,tau_a=1,tau_b=1,beta=c(0,0,0,0), b=c(0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),a=c(0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
 0,0,0,0,0,0,0,0,0),beta0=0)
```

Se se tiver mais do que uma cadeia deve-se criar uma lista do mesmo género para cada cadeia, com valores iniciais diferentes, colocando em objetos com nomes, por exemplo, Inits1,Inits2,... e dar a instrução

```
Inits<-list(Inits1,Inits2,...)
```

sendo portanto uma lista de listas.

6. A execução do software `OpenBUGS` pode ser agora feita através do `R` utilizando a função `bugs()`, depois de ter garantido que o pacote `OpenBUGS` se encontra carregado na sessão do `R`.

```
library(R2OpenBUGS)

Cexemplo1_openBUGS.fit<- bugs(data=Cexemplo1.data, inits=list(Inits),
parameters.to.save=Cexemplo1.params,
"Cexemplo1BUGS.txt", n.chains=1, n.iter=40000,
n.burnin=20000, debug=FALSE,save.history=FALSE,DIC=TRUE)
```

É conveniente averiguar quais as componentes que constituem a função `bugs()` através da instrução ?`bugs()`.

7. Para obter um sumário da distribuição *a posteriori* marginal dos parâmetros que foram declarados no vetor guardado no objeto `Cexemplo1.params`, escreve-se

```
Cexemplo1_OpenBUGS.fit$summary
```

obtendo-se neste caso

```
              mean        sd     2.5%       25%       50%       75%     97.5%
beta0      17.4290    1.9208  13.9200   16.1000   17.2800   18.6600   21.3800
beta[1]     4.2329    2.8568  -2.6160    2.5750    4.4480    6.0600    9.2410
```

```
beta[2]    0.1422   0.1391 -0.1303   0.0474   0.1448   0.2352   0.4132
beta[3]    4.3456   0.9455  2.3960   3.7600   4.3430   4.9100   6.3730
beta[4]   -0.1029   0.0366 -0.1726  -0.1281  -0.1046  -0.0798  -0.0241
tau        1.5607   3.6155  0.0636   0.0847   0.1521   1.1012  13.2300
tau_a      0.0162   0.0038  0.0097   0.0135   0.0159   0.0185   0.0243
tau_b      2.2989   5.4749  0.0638   0.0870   0.1604   1.4452  19.2502
sigma_a    8.0226   0.9547  6.4140   7.3560   7.9320   8.5990  10.1600
sigma_b    2.1790   1.2991  0.2279   0.8318   2.4965   3.3910   3.9600
sigma      2.2725   1.2624  0.2750   0.9528   2.5640   3.4360   3.9660
deviance 583.7390 241.1451 37.2990 403.3750 695.3000 782.2000 810.8000
```

O que aqui se obtém são medidas sumárias das distribuições *a posteriori* marginais dos parâmetros do modelo, como seja a média, desvio padrão e quantis de probabilidades 0.025, 0.25, 0.5, 0.75 e 0.975. Por exemplo, ao olhar para estes elementos pode dizer-se que uma estimativa de β_3 (coeficiente da covariável z3) é 4.3456 e um intervalo de credibilidade de 95% é (2.396,6.373). Para obter intervalos HPD tem de se recorrer ao `CODA` ou `BOA`. Ir-se-á ver na subsecção concernente a tais pacotes como estes intervalos podem ser obtidos.

8. Diversos outros elementos estão presentes em `Cexemplo1_OpenBUGS.fit`. Usando a instrução

```
names(Cexemplo1_OpenBUGS.fit)
```

obtém-se a seguinte lista dos elementos desse objeto:

```
 [1] "n.chains"         "n.iter"
 [3] "n.burnin"         "n.thin"
 [5] "n.keep"           "n.sims"
 [7] "sims.array"       "sims.list"
 [9] "sims.matrix"      "summary"
[11] "mean"             "sd"
[13] "median"           "root.short"
[15] "long.short"       "dimension.short"
[17] "indexes.short"    "last.values"
[19] "isDIC"            "DICbyR"
[21] "pD"               "DIC"
[23] "model.file"
```

Por exemplo, obtém-se informação sobre o DIC escrevendo

```
Cexemplo1_OpenBUGS.fit$DIC
[1] 429.7
Cexemplo1_OpenBUGS.fit$pD
[1] -154.1
```

O valor negativo de p_D pode significar excesso de parâmetros no modelo. Esta situação pode ser devida ao facto de se ter considerado no modelo os efeitos aleatórios "b".

9. Por fim é imprescindível fazer um estudo de convergência do processo com recurso aos métodos referidos no capítulo 6. Quer o pacote `CODA`, quer o pacote `BOA` do `R` podem ser usados para o efeito.

10. Uma análise completa do modelo irá envolver a seleção do melhor modelo e um estudo de adequabilidade. Todo esse estudo pode ser feito através do `R`. Para tal basta ter acesso aos valores simulados dos parâmetros do modelo. Esses valores encontram-se na forma de lista, *array* ou matriz. Esta última, por exemplo, obtém-se através da instrução

```
A<-Cexemplo1_OpenBUGS.fit$sims.matrix
dim(A)
[1] 20000    12
head(A)# mostra as seis primeiras linhas de A
       beta0 beta[1] beta[2] beta[3] beta[4]   tau
[1,] 15.45   8.683   0.066   4.086  -0.118 0.422
[2,] 16.44   5.370   0.123   4.398  -0.117 0.071
[3,] 19.35   2.387   0.214   4.675  -0.037 2.613
[4,] 19.12  -0.014   0.152   4.254  -0.084 0.361
[5,] 17.49   3.548   0.068   5.744  -0.104 2.452
[6,] 19.98   3.522   0.384   3.826  -0.106 0.092
     tau_a tau_b sigma_a sigma_b sigma deviance
[1,] 0.017 0.078   7.640   3.589 1.539    536.6
[2,] 0.016 2.567   7.895   0.624 3.749    790.8
[3,] 0.019 0.074   7.169   3.665 0.619    249.5
[4,] 0.016 0.107   7.853   3.061 1.665    570.3
[5,] 0.015 0.079   8.067   3.558 0.639    286.1
[6,] 0.020 3.179   7.058   0.561 3.298    791.9
```

De notar que só aparecem os valores simulados dos parâmetros que se declararam no objeto "Cexemplo1.params". Para fazer um estudo dos resíduos, por exemplo, é importante ter monitorizado o "mu".

Se se retirarem os efeitos aleatórios "b" do modelo (de notar que tem de se alterar o ficheiro *txt* que define o modelo e a lista com os valores iniciais) obtém-se

```
> exemplo1_OpenBUGS1.fit<- bugs(data=Cexemplo1.data, inits=list(Inits1),
 parameters.to.save=Cexemplo1.params1,
+ "Cexemplo1BUGS_semb.txt", n.chains=1,
n.iter=40000,n.burnin=20000, debug=FALSE,save.history=FALSE,DIC=TRUE)

> exemplo1_OpenBUGS1.fit$summary
             mean      sd     2.5%      25%      50%      75%    97.5%
beta0     17.2103  1.7634  13.7600  16.0400  17.1900  18.3600  20.7900
beta[1]    4.7809  2.3927   0.0629   3.1860   4.7305   6.4165   9.4950
beta[2]    0.1524  0.1447  -0.1171   0.0517   0.1492   0.2492   0.4498
beta[3]    4.1461  0.9225   2.3839   3.5270   4.1210   4.7510   6.0530
beta[4]   -0.1069  0.0360  -0.1774  -0.1309  -0.1073  -0.0826  -0.0357
tau        0.0773  0.0112   0.0570   0.0695   0.0768   0.0846   0.1008
tau_a      0.0164  0.0038   0.0100   0.0137   0.0160   0.0187   0.0247
sigma_a    7.9739  0.9355   6.3630   7.3190   7.8970   8.5440  10.0200
sigma      3.6252  0.2650   3.1500   3.4390   3.6080   3.7940   4.1900
deviance 795.1262 12.6237 772.6000 786.2000 794.3000 803.2000 822.0000

> exemplo1_OpenBUGS1.fit$DIC
[1] 843.2
> exemplo1_OpenBUGS1.fit$pD
[1] 48.03
> A1<-exemplo1_OpenBUGS1.fit$sims.matrix
> dim(A1)
[1] 20000    10
> head(A1)
      beta0 beta[1] beta[2] beta[3] beta[4]    tau
[1,] 20.62 -0.8067  0.1487   4.631 -0.1005 0.0838
[2,] 18.39  2.8580  0.0692   5.767 -0.0569 0.0826
[3,] 19.56  2.5330  0.3619   3.045 -0.1277 0.0784
[4,] 16.61  6.1660 -0.1078   4.284 -0.1168 0.0827
[5,] 18.11  2.8790  0.1680   4.372 -0.2055 0.0636
[6,] 15.06  5.1330  0.0461   3.707 -0.1142 0.0639
      tau_a sigma_a sigma deviance
[1,] 0.0188   7.291 3.455    773.9
[2,] 0.0141   8.431 3.479    781.4
```

```
[3,] 0.0247   6.368 3.572    796.0
[4,] 0.0162   7.844 3.478    783.5
[5,] 0.0251   6.310 3.964    819.5
[6,] 0.0168   7.704 3.955    818.8

> Cexemplo1_OpenBUGS1$DIC
[1] 843.1
> Cexemplo1_OpenBUGS1$pD
[1] 48.12
```

Note-se que p_D é agora positivo e igual a 48.12.

8.3 JAGS

Em 2003 Martyn Plummer, da *International Agency for Research on Cancer*, criou o `JAGS` (*Just Another Gibbs Sampler*), um clone do `BUGS` escrito em `C++`, o qual veio corrigir certos aspetos negativos do `BUGS`. Os modelos escritos em linguagem `BUGS` podem ser usados no `JAGS` praticamente sem alterações, tendo a vantagem de correr em todas as plataformas. Há duas partes na definição de um modelo em `JAGS`: a descrição do modelo (`model{}`), tal como no `BUGS`, e a definição dos dados (`data{}`). Este bloco pode ser usado, por exemplo, para definir transformações dos dados, definir estatísticas sumárias, simular conjuntos de dados, etc. A última versão do `JAGS` foi lançada em julho de 2017 (JAGS 4.3.0). A leitura do manual (Plummer, 2012; Plummer, 2017) é importante para perceber como o `JAGS` funciona. O pacote `R2jags` (https://cran.r-project.org/web/packages/R2jags/R2jags.pdf) serve como interface entre o `R` e o `JAGS`. Outra das grandes vantagens do `JAGS` em relação ao OpenBUGS é a rapidez de processamento.

8.3.1 Exemplo de aplicação: recurso ao `R2jags`

No que se segue ilustra-se como se pode usar o `R2jags` para estudar o modelo do exemplo em 8.1 desprovido dos efeitos aleatórios b. O software pode ser instalado e carregado respetivamente através das instruções

```
install.packages("R2jags", dependencies=TRUE,
repos="http://cran.us.r-project.org")
```

```
library(R2jags)
```

1. O mesmo modelo usado no OpenBUGS pode ser dado como ficheiro de texto, ou pode ser definido, no *script* do R, como uma função, do seguinte modo:

```
exemplo1.model<-function(){
for(i in 1:147){
X[i]~dnorm(mu[i],tau)
mu[i]<-beta0+beta[1]*z1[i]+beta[2]*z2[i]+beta[3]*z3[i]
+beta[4]*z[i]+a[ID[i]]
}
for(j in 1:49){
a[j]~dnorm(0,tau_a)
}
for(k in 1:4){
beta[k]~dnorm(0,0.0001)
}
beta0~dnorm(0,0.0001)
tau~dgamma(0.05,0.05)
tau_a~dgamma(0.05,0.05)
sigma<-1/sqrt(tau)
sigma_a<-1/sqrt(tau_a)
}
```

2. A leitura dos dados, a definição das variáveis que entram no modelo, a definição dos parâmetros a monitorizar e dos valores iniciais dos parâmetros, faz-se exatamente do mesmo modo que anteriormente.

```
Cexemplo1.data <- list("X","ID","z3","z1","z2","z")
Cexemplo1.params1 <- c("beta0","beta","tau","tau_a",
 "sigma_a","sigma")

Inits1<-list("tau"=1,"tau_a"=1,"beta"=c(0,0,0,0),
"a"=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
"beta0"=0)
```

3. Antes de usar o R2jags pela primeira vez pode haver necessidade de definir uma semente. Para isso pode, por exemplo, escrever-se no R a instrução

```
set.seed(123)
```

4. Para adaptar o modelo no JAGS usa-se a função `jags()`.

```
exemplo1_JAGS.fit <- jags(data = Cexemplo1.data, inits = list(Inits1),
parameters.to.save = Cexemplo1.params1, n.chains = 1, n.iter = 40000,
n.burnin = 20000, model.file = exemplo1.model)
```

5. Estatísticas sumárias relativas às distribuições *a posteriori* marginal dos parâmetros, obtêm-se através da instrução

```
> print(exemplo1_JAGS.fit)
Inference for Bugs model at "C:/Users...model1f5469ec52a3.txt",
fit using jags,
 1 chains, each with 40000 iterations (first 20000 discarded),
  n.thin = 20, n.sims = 1000 iterations saved
         mu.vect sd.vect    2.5%     25%     50%     75%   97.5%
beta[1]    4.751   2.537  -0.248   3.038   4.783   6.507   9.450
beta[2]    0.160   0.143  -0.121   0.069   0.159   0.249   0.447
beta[3]    4.163   0.915   2.467   3.552   4.132   4.777   6.081
beta[4]   -0.107   0.036  -0.177  -0.132  -0.108  -0.081  -0.031
beta0     17.212   1.810  13.561  16.050  17.151  18.387  20.717
sigma      3.676   0.808   3.133   3.441   3.604   3.789   4.267
sigma_a    7.920   1.159   6.289   7.335   7.921   8.529   9.778
tau        0.077   0.013   0.055   0.070   0.077   0.084   0.102
tau_a      0.699  13.406   0.010   0.014   0.016   0.019   0.025
deviance 797.267  28.396 773.207 786.080 794.391 802.998 823.233

DIC info (using the rule, pD = var(deviance)/2)
pD = 403.2 and DIC = 1200.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

Note-se o intervalo entre simulações (n.thin=20). De facto uma das componentes da função `jags()` é precisamente `n.thin`. Caso não seja definido pelo utilizador é usado o valor predefinido dado por `max(1, floor((n.iter - n.burnin)/1000))`; atente-se também no valor elevado de pD.

6. Uma outra vantagem do `R2jags` é a possibilidade de "deixar ao cuidado" do programa a geração de valores iniciais das cadeias escrevendo `NULL` no lugar dos valores iniciais. Vai-se exemplificar isto com duas cadeias.

```
exemplo1_JAGS.fit2 <- jags(data = Cexemplo1.data, inits = NULL,
```

```
parameters.to.save = Cexemplo1.params1, n.chains = 2, n.iter = 40000,
n.burnin = 20000, model.file = exemplo1.model)

> print(exemplo1_JAGS.fit2)
Inference for Bugs model at
"C:/Users/TOSHIBA1/AppData/Local/Temp/RtmpkpwgcN/model1f5477c03ee1.txt",
 fit using jags,
 2 chains, each with 40000 iterations (first 20000 discarded),
 n.thin = 20
 n.sims = 2000 iterations saved
         mu.vect sd.vect    2.5%     25%     50%     75%   97.5%
beta[1]    4.859   2.596  -0.346   3.166   4.939   6.604   9.950
beta[2]    0.157   0.139  -0.113   0.068   0.151   0.242   0.453
beta[3]    4.191   0.918   2.422   3.595   4.173   4.805   6.059
beta[4]   -0.105   0.036  -0.175  -0.130  -0.106  -0.081  -0.035
beta0     17.218   1.816  13.857  15.996  17.207  18.435  20.716
sigma      3.659   0.715   3.164   3.436   3.608   3.798   4.235
sigma_a    7.943   1.049   6.363   7.317   7.886   8.587   9.873
tau        0.077   0.012   0.056   0.069   0.077   0.085   0.100
tau_a      0.507  14.898   0.010   0.014   0.016   0.019   0.025
deviance 796.175  22.494 773.690 786.454 794.169 802.528 822.873
          Rhat n.eff
beta[1]  1.001  2000
beta[2]  1.003   590
beta[3]  1.001  2000
beta[4]  1.001  2000
beta0    1.001  2000
sigma    1.006  2000
sigma_a  1.040  2000
tau      1.006  2000
tau_a    1.040  2000
deviance 1.022  2000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 253.1 and DIC = 1049.3
DIC is an estimate of expected predictive error (lower deviance is better)
```

Note-se que há agora duas colunas extra "Rhat" e "n.eff", cujo significado está discriminado no *output* do programa.

Um gráfico dos valores simulados dos parâmetros pode ser obtido através da instrução

```
traceplot(exemplo1_JAGS.fit2)
```

7. Estes gráficos podem indicar de imediato se há indícios de não convergência. No caso de se verificar que não há convergência pode continuar-se a simulação até convergência, através da instrução (que só se pode usar quando há pelo menos duas cadeias)

```
exemplo1_JAGS.fit2.upd <- autojags(exemplo1_JAGS.fit2)
```

```
print(exemplo1_JAGS.fit2.upd)
Inference for Bugs model at
"C:/Users/TOSHIBA1/AppData/Local/.../model1f5477c03ee1.txt",
 fit using jags,
 2 chains, each with 1000 iterations (first 0 discarded)
 n.sims = 2000 iterations saved
         mu.vect sd.vect    2.5%     25%     50%     75%   97.5%
beta[1]    4.681   2.562  -0.227   2.972   4.702   6.345   9.832
beta[2]    0.156   0.143  -0.117   0.059   0.151   0.253   0.440
beta[3]    4.179   0.924   2.410   3.575   4.181   4.771   5.951
beta[4]   -0.106   0.037  -0.178  -0.132  -0.106  -0.081  -0.036
beta0     17.288   1.777  13.735  16.117  17.333  18.443  20.676
sigma      3.631   0.265   3.139   3.444   3.611   3.801   4.210
sigma_a    8.007   0.955   6.314   7.363   7.936   8.582  10.080
tau        0.077   0.011   0.056   0.069   0.077   0.084   0.102
tau_a      0.016   0.004   0.010   0.014   0.016   0.018   0.025
deviance 795.438  12.641 773.639 786.407 794.608 802.980 823.357
          Rhat n.eff
beta[1]  1.001  2000
beta[2]  1.001  2000
beta[3]  1.001  2000
beta[4]  1.001  2000
beta0    1.001  2000
sigma    1.002  2000
sigma_a  1.001  2000
tau      1.002  2000
```

```
tau_a    1.001  2000
deviance 1.002  2000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 79.9 and DIC = 875.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

Note-se a redução dos valores de p_D e `DIC`. Outra função que também se pode usar para o efeito e que exige apenas uma cadeia é a função `update()`.

Numa secção posterior faremos um estudo completo deste exemplo, indicando ainda outros pacotes que podem ser usados para fazer estudo da convergência.

8.4 Stan

Em 2010, Andrew Gelman da Columbia University em Nova Iorque, e colaboradores, estavam a trabalhar na análise bayesiana de modelos lineares generalizados em múltiplos níveis descritos em Gelman e Hill (2007). A implementação destes modelos no `WinBUGS` ou `JAGS` mostrou ser extremamente desafiante devido à sua complexa estrutura. Por exemplo, Matt Schofield verificou que o modelo de séries temporais multinível que estava a utilizar no estudo da reconstrução do clima, usando medições dos anéis de árvores, não convergia depois de centenas de milhares de iterações (Schofield *et al.*, 2014). Para resolver este problema Gelman e colaboradores desenvolveram um novo software bayesiano, que batizaram de `Stan`, em homenagem a Stanislaw Ulam, um dos criadores do método de Monte-Carlo. A primeira versão foi posta à disposição dos utilizadores em Agosto de 2012. O `Stan` não recorre ao amostrador de Gibbs mas sim a uma variante de Monte Carlo hamiltoniano (Neal, 2011). Com o algoritmo por eles desenvolvido, o *no-U-turn sampler* (Hoffman and Gelman, 2011, 2014), todos os parâmetros de um modelo são simulados em bloco e, desta forma, os problemas de convergência são substancialmente atenuados. Contrariamente ao `BUGS`, `Stan` está escrito numa linguagem imperativa.

Stan permite a utilização de todos os operadores básicos e funções da linguagem C++, para além de um grande número de outras funções especiais nomeadamente, funções de Bessel, funções gama e digama, e uma variedade de funções de ligação e suas inversas, utilizadas nos modelos lineares generalizados. Uma lista completa dos operadores básicos e funções especiais implementadas no Stan encontra-se em Carpenter *et al.* (2015). No que diz respeito a distribuições de probabilidade a lista, que se encontra ainda nesse artigo, também é vasta, o que permite uma grande flexibilidade na construção de modelos.

Um pouco da história do desenvolvimento do Stan, assim como detalhes da sua implementação, pode ser lida *no Stan Modeling Language: User's Guide and Reference Manual* correspondente à versão 2.6.2. A ligação entre o R e o **Stan** é feita através do RStan, (Stan Development Team, 2014 b).

8.4.1 Exemplo de aplicação: recurso ao RStan

Para instalar o RStan tem de se seguir as instruções que se encontram em https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started.

1. Tal como nos pacotes anteriores tem de se definir o modelo em linguagem Stan. Este modelo deve ser guardado num ficheiro de texto, tipicamente com um sufixo .stan. A definição do modelo aqui é um pouco mais elaborada do que nos pacotes anteriores. Sendo uma linguagem imperativa, é importante a ordem pelas quais são dadas as instruções. Para o exemplo que estamos vindo a desenvolver, a definição do modelo para ser lido na linguagem Stan e guardado num ficheiro "exemplo1.stan.txt", é como segue:

```
data {
int<lower=1> J; // length of data
int<lower=1> N; // number of patients
real X[J]; // response variable
real z2[J]; // covariate
real z3[J]; // covariate
real z[J]; // covariate
int<lower=0,upper=1> z1[J]; // z1 takes values 0 and 1
int<lower=1> ID[J]; // identification
}
parameters {
real<lower=0> tau;
```

```
real<lower=0> tau_a;
real beta0;
real beta[4];
real a[N];
}
transformed parameters {
real<lower=0> sigma_a;
real<lower=0> sigma;
sigma_a<-sqrt(1/tau_a);
sigma<-sqrt(1/tau);
}
model {
vector[J] mu;
for(i in 1:J){
mu[i] <-beta0+beta[1]*z1[i]+beta[2]*z2[i]
+beta[3]*z3[i]+a[ID[i]]+beta[4]*z[i]};
beta0 ~ normal(0,100);
beta ~ normal(0,100);
tau ~ gamma(0.05,0.05);
tau_a ~ gamma(0.05,0.05);
a ~ normal(0, sigma_a);
X ~ normal(mu, sigma);
}
generated quantities {
vector[J] log_lik;
vector[J] m;
for(i in 1:J) {
m[i] <-beta0+beta[1]*z1[i]+beta[2]*z2[i]
+beta[3]*z3[i]+a[ID[i]]+beta[4]*z[i]};
log_lik[i] <- normal_log(X[i], m[i], sigma);
}
}
```

Explicando este programa bloco por bloco, tem-se que:

- No bloco `data` devem ser declarados os dados a serem utilizados durante a execução do `Stan`. Por exemplo, tanto J como N são inteiros sendo o menor valor que tomam 1; `z1` sendo uma covariável binária é um inteiro cujo menor valor é 0 e o maior é 1. A variável resposta e as covariáveis são vetores reais. `Stan` admite ainda dados na forma de matrizes, vetores

ordenados, *arrays*, etc.

- No bloco `parameters` incluem-se todas as quantidades desconhecidas que o `Stan` irá estimar. Neste caso decidiu-se incluir as precisões `tau` e `tau_a` para estar de acordo com o modelo definido anteriormente. Novamente aqui define-se o tipo de quantidade que se quer estimar. Por exemplo `tau` e `tau_a` são reais positivos; `beta` e `a` são vetores reais de dimensão `4` e `N`, respetivamente.

- No bloco `transformed parameters` devem ser definidas as quantidades que são funções dos dados e/ou dos parâmetros. Neste caso definiu-se `sigma` e `sigma_a` como sendo a raíz quadrada do inverso da precisão respetiva.

- No bloco `model` define-se o modelo propriamente dito. Define-se o vetor `mu` em função das covariáveis e dos efeitos aleatórios `a`. Define-se o modelo para a variável resposta (note-se que aqui o segundo parâmetro da distribuição normal é o desvio padrão). Definiram-se também as distribuições *a priori* dos parâmetros e hiperparâmetros do modelo. Não é obrigatório no `Stan` definir as distribuições *a priori*. Na ausência destas definições o `Stan` procede usando distribuições *a priori* uniformes.

- O bloco `generated quantities` não é obrigatório. Definiu-se aqui a log-verosimilhança de modo a que os termos individuais possam ser guardados pelo Stan. O código está formulado de modo que estes termos sejam coligidos num objeto com o nome "log-lik"; os valores simulados dos elementos deste objeto são usados para calcular o WAIC e o LOO (este último é outra medida de desempenho, *vide* Vehtari and Gelman, 2014), como se verá.

2. Os dados podem ser introduzidos como uma lista, como anteriormente, ou se os dados forem lidos de um ficheiro, basta criar um objeto com os nomes das variáveis, tal como segue.

```
Cexemplo1<-read.table("Cexemplo1.txt",header=T)
attach(Cexemplo1)
J<-nrow(Cexemplo1) #J=147
N<-length(unique(ID)) #N=49
# objeto a ser usado pelo Stan
Cexemplo1_data<-c("N","J","X","ID","z3","z2","z1","z")
```

3. Seguidamente pode-se chamar a função `stan()` da programoteca `rstan` para simular da distribuição *a posteriori*:

```
library(rstan)
exemplo1.fit_stan <- stan(file="exemplo1.stan.txt",
 data=Cexemplo1_data, iter=40000, chains=2)
```

Veja-se usando `?stan` os argumentos da função stan. Particularmente úteis são os argumentos `sample_file` e `diagnostic_file` que permitem indicar os nomes dos ficheiros onde se pretende guardar as amostras simuladas de todos os parâmetros do modelo e dados para diagnóstico de convergência, respetivamente. Se estes nomes não forem indicados esses elementos não serão guardados. Podem, no entanto, ser extraídos posteriormente.

O `rstan` dá informação sobre o tempo de execução de cada cadeia na seguinte forma (apresenta-se aqui só para uma cadeia):

```
 COMPILING THE C++ CODE FOR MODEL 'example1' NOW.

SAMPLING FOR MODEL 'exempl1' NOW (CHAIN 1).

Chain 1, Iteration:     1 / 40000 [  0%]  (Warmup)
Chain 1, Iteration:  4000 / 40000 [ 10%]  (Warmup)
Chain 1, Iteration:  8000 / 40000 [ 20%]  (Warmup)
Chain 1, Iteration: 12000 / 40000 [ 30%]  (Warmup)
Chain 1, Iteration: 16000 / 40000 [ 40%]  (Warmup)
Chain 1, Iteration: 20000 / 40000 [ 50%]  (Warmup)
Chain 1, Iteration: 20001 / 40000 [ 50%]  (Sampling)
Chain 1, Iteration: 24000 / 40000 [ 60%]  (Sampling)
Chain 1, Iteration: 28000 / 40000 [ 70%]  (Sampling)
Chain 1, Iteration: 32000 / 40000 [ 80%]  (Sampling)
Chain 1, Iteration: 36000 / 40000 [ 90%]  (Sampling)
Chain 1, Iteration: 40000 / 40000 [100%]  (Sampling)
#  Elapsed Time: 46.431 seconds (Warm-up)
#                75.594 seconds (Sampling)
#                122.025 seconds (Total)
```

O tempo de execução de 40000 iterações é muito maior do que o tempo de execução do `OpenBUGS` ou do `JAGS`, o qual para este problema é praticamente

inexistente. Contudo, o número de iterações necessárias para atingir convergência usando o `Stan` é inferior. Aqui optou-se por usar o mesmo número de iterações mas não havia necessidade. Com efeito, se não se indicar o número de iteradas em `iter=`, observa-se neste exemplo convergência ao fim de 2000 iteradas.

4. Para obter estatísticas sumárias das distribuições *a posteriori* marginais dos parâmetros de interesse usa-se a instrução:

```
print(exemplo1.fit_stan,
pars=c("beta0","beta","sigma","sigma_a","tau","tau_a","lp__"))
```

obtendo-se

```
Inference for Stan model: example1.
2 chains, each with iter=40000; warmup=20000; thin=1;
post-warmup draws per chain=20000, total post-warmup draws=40000.

           mean se_mean   sd    2.5%     25%     50%     75%   97.5%
beta0     17.22    0.02 1.79   13.67   16.04   17.23   18.41   20.72
beta[1]    4.80    0.03 2.52   -0.17    3.12    4.80    6.48    9.77
beta[2]    0.16    0.00 0.14   -0.12    0.06    0.16    0.25    0.44
beta[3]    4.19    0.01 0.93    2.38    3.55    4.18    4.82    6.00
beta[4]   -0.10    0.00 0.04   -0.18   -0.13   -0.11   -0.08   -0.03
sigma      3.62    0.00 0.26    3.16    3.44    3.61    3.79    4.19
sigma_a    8.00    0.01 0.94    6.39    7.34    7.92    8.57   10.08
tau        0.08    0.00 0.01    0.06    0.07    0.08    0.08    0.10
tau_a      0.02    0.00 0.00    0.01    0.01    0.02    0.02    0.02
lp__    -388.89    0.06 6.36 -402.47 -392.90 -388.50 -384.42 -377.55
        n_eff Rhat
beta0    5771    1
beta[1]  5754    1
beta[2]  7261    1
beta[3]  7555    1
beta[4] 10087    1
sigma   21771    1
sigma_a 23821    1
tau     21546    1
tau_a   25151    1
lp__    11599    1
```

```
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Aqui `se_mean` é o erro padrão de Monte Carlo da média. Caso não tivesse sido indicado quais os parâmetros para os quais se pretende estatísticas sumárias, essas estatísticas seriam apresentadas para as distribuições *a posteriori* marginais de todas as quantidades desconhecidas, nomeadamente também para `a, log_lik, m`. A quantidade `lp_` que aparece como último elemento em `pars=` e cujas estatísticas sumárias se encontram na última linha, é o logaritmo da densidade *a posteriori* (não normalizada) calculada pelo `Stan` durante a execução do algoritmo de Monte Carlo hamiltoniano. Esta quantidade pode ser usada na avaliação e comparação de modelos (ver, por exemplo, Vehtari and Ojanen, 2012).

Com a função `print()` obtém-se estatísticas sumárias juntando todas as cadeias que se geraram. Com a função `summary()` obtém-se as estatísticas sumárias para cada cadeia em separado.

5. Para obter os valores simulados dos parâmetros usa-se a função `extract()`. Quando se coloca no argumento `permute=TRUE` é criada uma lista com os valores simulados de todos os parâmetros (caso se queira apenas valores simulados de alguns dos parâmetros, tal deve ser indicado no argumento `pars=`); `se permute=FALSE`, é criado um *array* cuja primeira dimensão corresponde às iteradas, a segunda corresponde às cadeias e a terceira aos parâmetros. Veja-se um exemplo:

```
#####usando permuted=TRUE###############
samples_stan<-extract(exemplo1.fit_stan,
pars=c("beta0", "beta", "sigma", "sigma_a"),
permuted = TRUE, inc_warmup = FALSE, include = TRUE)
> class(samples_stan)
[1] "list"
> names(samples_stan)
[1] "beta0"   "beta"    "sigma"   "sigma_a"
> length(samples_stan$beta0)
[1] 40000 #20000 para cada cadeia
> head(samples_stan$beta0)
[1] 16.18767 18.64417 20.43510 16.69809 14.35278 15.39996
> dim(samples_stan$beta)
```

```
> head(round(samples_stan$beta,3))

iterations   [,1]  [,2]  [,3]   [,4]
      [1,]  8.994 0.468 2.437 -0.126
      [2,]  4.310 0.309 4.425 -0.093
      [3,]  2.127 0.079 3.700 -0.156
      [4,]  3.394 0.245 1.680 -0.131
      [5,] 10.541 0.359 3.814 -0.086
      [6,]  8.314 0.357 4.001 -0.068
#########usando permuted=FALSE#########
> samples_stan_array<-extract(exemplo1.fit_stan,
+ pars=c("beta0", "beta", "sigma", "sigma_a"),
+ permuted = FALSE, inc_warmup = FALSE, include = TRUE)
> class(samples_stan_array)
[1] "array"
> dim(samples_stan_array)
[1] 20000     2     7 #20000 cada cadeia, 2 cadeias, 7 parâmetros
> samples_stan_array[1:4,1:2,1:3]
, , parameters = beta0

          chains
iterations  chain:1  chain:2
      [1,] 16.29099 17.81893
      [2,] 16.68243 17.31063
      [3,] 16.49383 17.31063
      [4,] 16.20388 16.70740

, , parameters = beta[1]

          chains
iterations  chain:1  chain:2
      [1,] 6.530125 5.718621
      [2,] 4.949012 6.479835
      [3,] 6.000288 6.479835
      [4,] 6.204705 7.421142

, , parameters = beta[2]

          chains
iterations   chain:1    chain:2
```

```
[1,] 0.1718956 0.07575568
[2,] 0.1657402 0.18286167
[3,] 0.1793824 0.18286167
[4,] 0.1347633 0.15160846
```

6. Com a função `traceplot()` obtém-se um gráfico dos valores simulados dos parâmetros para cada cadeia. A Figura 8.1 resulta da seguinte instrução:

```
traceplot(exemplo1.fit_stan,
pars=c("beta"), nrow = 5, ncol = 2, inc_warmup = FALSE)
```

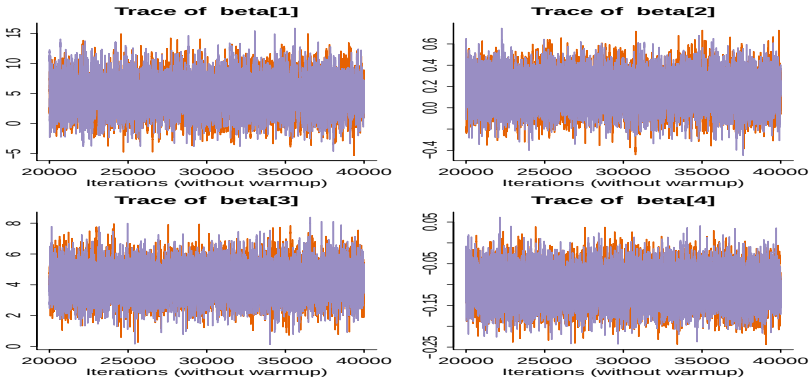

Figura 8.1: Gráfico dos valores simulados

7. Para obter o WAIC (ver Vehtari *et al.*, 2015) procede-se do seguinte modo, depois de se ter instalado o programa `loo` através do CRAN:

```
> library(loo)
> log_lik1 <- extract_log_lik(exemplo1.fit_stan)
> waic(log_lik1)
Computed from 40000 by 147 log-likelihood matrix

          Estimate   SE
elpd_waic   -422.4  9.9
p_waic        40.5  4.7
waic         844.8 19.8
```

Mais detalhes sobre como utilizar o `Rstan` pode encontrar-se em https://cran.r-project.org/web/packages/rstan/vignettes/rstan_vignette.pdf.

8.5 BayesX

BayesX é um software desenvolvido no Departamento de Estatística da Universidade de Munique por Andreas Brezger, Thomas Kneib e Stefan Lang, tendo as primeiras versões surgido em 2002. Este software foi criado especificamente para estimar modelos de regressão aditiva estruturada (Brezger *et al.*, 2005). Esta família de modelos (STAR), já referida no capítulo 7, engloba vários modelos de regressão bem conhecidos e muito usados nas aplicações, como sejam, os modelos aditivos generalizados (GAM), modelos mistos aditivos generalizados (GAMM), modelos mistos geoaditivos generalizados (GGAMM), modelos dinâmicos, modelos espaço-temporais, etc, através de uma estrutura unificada (Umlauf *et al.*, 2015). O BayesX, escrito em linguagem C++, permite a análise de modelos de regressão em que a variável resposta não pertence necessariamente à família exponencial. Permite ainda a análise de regressão quantílica, análise de sobrevivência através da modelação da função *hazard* (extensões do modelo de Cox), análise de modelos com múltiplos estados e de modelos com múltiplos níveis. O manual metodológico do BayesX (http://www.statistik.lmu.de/ bayesx/manual/methodology_manual.pdf) contém uma breve descrição da metodologia dos modelos de regressão admitidos pelo BayesX.

Os argumentos específicos da função do BayesX para a implementação de um modelo estatístico compreendem a especificação da família de distribuições da variável resposta, o método des estimação e outros parâmetros de controlo definidos através da função `bayesx.control()` como se verá. Tal como nos modelos `glm()` do R, a família de distribuições fixada por omissão é a gaussiana. Para além das distribuições habituais da família exponencial, uma lista de distribuições de probabilidade admitidas pelo BayesX encontra-se em Umlauf *et al.* (2015). Uma outra particularidade do BayesX é que, para além do método de estimação MCMC, o BayesX permite fazer um estudo inferencial dos modelos mistos usando o método de máxima verosimilhança restrita (REML) e o método da verosimilhança penalizada (STEP).

Em princípio é possível usar, quer o WinBUGS/OpenBUGS, quer o JAGS, para implementar os modelos STAR de eleição do BayesX. Os autores deste software afirmam que, em relação ao WinBUGS/OpenBUGS, há uma redução substancial no tempo de execução do programa, para além de as cadeias de Markov resultantes convergirem mais rapidamente (Brezger *et al.*, 2005), apresentando

melhores propriedades de mistura. Para facilitar o uso e subsequente análise dos resultados emanados do BayesX, Kneib *et al.* (2014) construíram um pacote do R, também designado por BayesX, o qual permite ler e processar os resultados provenientes do BayesX. Contudo, com este pacote, os utilizadores ainda têm de ler os dados, adaptar os modelos e obter os ficheiros com os resultados, através do BayesX. Para aliviar esta tarefa Umlauf *et al.* (2015) introduziram um novo pacote do R, o R2BayesX, o qual faz a interface completa entre o R e o BayesX.

O manual do R2BayesX encontra-se em
https://cran.r-project.org/web/packages/R2BayesX/R2BayesX.pdf.

Para instalar o R2BayesX basta fazê-lo como habitualmente através da instrução

```
install.packages("R2BayesX", dependencies=TRUE,
repos="http://cran.us.r-project.org")
```

8.5.1 Exemplo de aplicação: recurso ao R2BayesX

A sintaxe utilizada pelo R2BayesX para implementar um modelo bayesiano é em tudo muito semelhante à sintaxe utilizada pelo R na implementação de modelos estatísticos.

1. Para implementar o modelo que se tem vindo a analisar basta, após a leitura dos dados, escrever a formula do modelo e chamar a função bayesx().

```
Cexemplo1<-read.table("Cexemplo1.txt",header=T)
library(R2BayesX)
modelo2_BayesX<-X~z2+z3+z1+z+sx(ID,bs="re")
exemplo1_BayesX<-bayesx(formula = modelo2_BayesX, data = Cexemplo1,
family = "gaussian", method = "MCMC",chains=2, seed = c(123,456),
control = bayesx.control(model.name = "bayes2x.estim",
outfile='C:/...', iterations = 40000L,
burnin = 20000L,dir.rm=T))
```

Passando a explicar o código acima:

- A fórmula colocada no objeto denominado modelo2_BayesX é a fórmula do preditor linear relativo à média da variável resposta. Tal como se

formulou de ínício o modelo em questão, os efeitos fixos `z2, z3, z1, z` entram linearmente no preditor linear. O `BayesX` é contudo particularmente útil para modelar efeitos não lineares das covariáveis através da utilização da função `sx()`. Por exemplo, se houvesse razão para crer que a covariável `z2` teria um efeito não linear, e se quisesse usar um *P-spline* para modelar esse efeito não linear, então `z2` deveria entrar no modelo através da especificação `sx(z2,bs="ps")`, em que no argumento `bs` se escolhe o tipo de base do termo. Desse mesmo modo são especificados os efeitos aleatórios. Veja-se como o efeito aleatório "a" é introduzido na fórmula através da função `sx()`. Lembre-se de que estes efeitos aleatórios dizem respeito ao indivíduo, caraterizado pelo seu número de identificação `ID`, e por essa razão é que o primeiro argumento da função sx() é ID. Para especificar que os efeitos aleatórios são i.i.d. normais de valor médio nulo e desvio padrão σ_a, escreve-se `bs="re"`. Para especificar o modelo com os efeitos aleatórios "b", tratado inicialmente, escrever-se-ia também `sx(all,bs="re")` (recorde-se que a variável `all` continha os inteiros de 1 a 147). Na tabela 4 de Umlauf *et al.* (2015) encontra-se uma lista de todas os diferentes argumentos admitidos pelo `BayesX` para o tipo de base `bs`.

- A função `bayesx()` adapta o modelo especificado. Esta função tem diversos argumentos muitos dos quais estão prefixados. Se não forem especificados outros, são esses os usados. Os únicos obrigatórios a introduzir são os dois primeiros: `formula` e `data`. Por omissão (*default*) a família é a gaussiana, o método é MCMC [34], o número de iteradas é 12000, o período de aquecimento (`burn-in`) é de 2000 e o intervalo entre iteradas (*step*) é 10 e o número de cadeias é 1. Estes valores, assim como outros aspetos, podem ser alterados no argumento `control` através da função `bayesx.control`; como de costume, para saber quais os argumentos desta função e o seu significado, pode usar-se a ajuda do `R`, dando a instrução `?bayesx.control`. O primeiro argumento desta função é o `model.name` onde se introduz o nome do modelo que vai conter os ficheiros dos resultados da execução do modelo através do `BayesX`; esses ficheiros irão ser guardados na diretoria especificada em `outfile`.

2. Utilizando aquele código obtém-se o seguinte resultado das estatísticas

[34] O BayesX implementa também o método REML e STEP; veja-se detalhes em Umlauf *et al.* (2015).

sumárias das distribuições marginais dos parâmetros:

```
> summary(exemplo1_BayesX)
### Chain_1
Call:
bayesx(formula = formula, data = data, weights = weights,
subset = subset, offset = offset, na.action = na.action,
contrasts = contrasts, control = control, model = model,
chains = NULL, cores = NULL)

Fixed effects estimation results:

Parametric coefficients:
               Mean      Sd    2.5%     50%   97.5%
(Intercept) 17.2313  1.8036 13.6702 17.2744 20.7688
z2           0.1557  0.1371 -0.1174  0.1566  0.4250
z3           4.2146  0.9665  2.3040  4.2043  6.2029
z1           4.7691  2.4910 -0.1460  4.7646  9.6957
z           -0.1043  0.0366 -0.1756 -0.1055 -0.0319

Random effects variances:
            Mean     Sd   2.5%    50%  97.5%    Min   Max
sx(ID):re 64.925 15.702 41.081 62.706 99.701 26.824 169.6

Scale estimate:
          Mean      Sd    2.5%     50%  97.5%
Sigma2 13.2389  1.9332  9.9936 13.0804 17.373

N = 147  burnin = 20000  DIC = 194.7823  pd = 48.29928
method = MCMC  family = gaussian  iterations = 40000  step = 10
### Chain_2
Call:
bayesx(formula=formula, data=data, weights=weights, subset=subset,
    offset = offset, na.action = na.action, contrasts = contrasts,
    control = control, model = model, chains = NULL, cores = NULL)

Fixed effects estimation results:

Parametric coefficients:
               Mean      Sd    2.5%     50%   97.5%
```

```
(Intercept) 17.1458  1.7125 13.9773 17.0971 20.3820
z2           0.1591  0.1438 -0.1282  0.1612  0.4407
z3           4.1544  0.9413  2.3008  4.1405  6.0312
z1           4.9990  2.5100 -0.2337  5.0116  9.6973
z           -0.1025  0.0351 -0.1751 -0.1016 -0.0367

Random effects variances:
             Mean      Sd    2.5%     50%   97.5%     Min    Max
sx(ID):re  64.569  15.502  40.542  62.367 101.027  30.008 136.28

Scale estimate:
          Mean      Sd    2.5%     50%  97.5%
Sigma2 13.2323  1.9739  9.9179 13.0191 17.518

N = 147  burnin = 20000  DIC = 195.1921  pd = 48.54314
method = MCMC  family = gaussian  iterations = 40000  step = 10
###
Object consists of 2 models
```

Como no argumento `control` se indicou um nome para o ficheiro com os resultados, são automaticamente criadas na diretoria especificada pastas, uma para cada cadeia, contendo diversos ficheiros com as amostras simuladas dos parâmetros (com a extensão .raw) e ficheiros de texto com estatísticas sumárias. No caso vertente foram criadas duas pastas com os nomes `Chain_1_bayes2x.estim` e `Chain_2_bayes2x.estim`. Todos estes ficheiros de dados podem ser utilizados depois para fazer representações gráficas, estudos de diagnóstico, etc.

3. A instrução

```
getscript(exemplo1_BayesX)
```

tem como resultado um *script* do `R` para fazer representações gráficas usando esses ficheiros.

4. Alternativamente, amostras dos valores simulados das distribuições *a posteriori* dos parâmetros, podem também ser obtidas através da função `samples()` do seguinte modo:

```
AA<-samples(exemplo1_BayesX)
```

```
> class(AA)
[1] "mcmc.list"
> names(AA)
[1] "Chain_1" "Chain_2"
> length(AA[[1]])
[1] 10000
> length(AA[[2]])
[1] 10000

plot(AA)
```

O objeto é uma lista contendo em AA[[1]] as amostras da distribuição *a posteriori* dos parâmetros dos efeitos fixos relativas à cadeia 1 e em AA[[2]] as relativas à cadeia 2. Como se guardaram 20000 iteradas com intervalo entre iteradas de 10, há 2000 valores simulados para cada um dos 5 parâmetros. A ordem por que os valores simulados aparecem é a ordem por que foram introduzidos na fórmula; assim AA[[1]][1:2000] contém valores simulados da distribuição *a posteriori* da ordenada na origem (`beta0`); AA[[1]][2001:4000] contém valores simulados da distribuição *a posteriori* de beta[1], coeficiente da covariável `z2`, etc.

5. A instrução `plot(AA)` faz um gráfico das séries sobrepostas dos valores simulados para cada parâmetro e um gráfico da densidade *a posteriori* marginal correspondente.

6. Para se obter valores simulados da distribuição *a posteriori* das variâncias (neste caso a variância da gaussiana), e respetivas representações gráficas, escreve-se:

```
> Va<-samples(exemplo1_BayesX,term = "var-samples")
> length(Va[[1]])
[1] 2000
> plot(Va)
```

Na Figura 8.2 aparece o resultado relativo à instrução `plot(Va)`.

7. Para se obter valores simulados dos efeitos aleatórios "a" e de σ_a^2 basta escrever no argumento `term` da função `samples term="sx(ID)"`. Do mesmo modo podemos obter amostras da distribuição *a posteriori* para parâmetros específicos; por exemplo, para o parâmetro correspondente a `z3` escreve-se `term="z3"`.

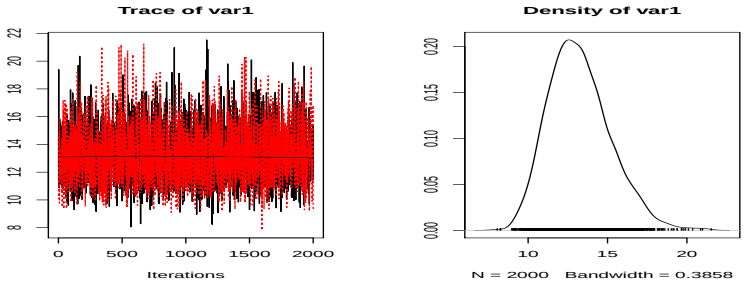

Figura 8.2: Traçado dos valores simulados e densidade da distribuição *a posteriori* marginal de σ^2

8. Um aspeto interessante é que, nas pastas criadas com os resultados das cadeias, aparece também um documento em `latex` com o sumário do modelo, incluindo informação sobre as distribuições *a priori* utilizadas e um sumário dos resultados obtidos. Por exemplo, por consulta desse documento e já que não se especificou as distribuições *a priori* a utilizar, fica-se a saber que foram usados *diffuse priors* para os efeitos fixos, efeitos aleatórios i.i.d. gaussianos e distribuições *a priori* gamma inversa com hiperparâmetros *a=0.001* e *b=0.001* para as componentes da variância. A especificação dos hiperparâmetros pode ser alterada no argumento `control`.

A informação contida neste documento em `latex` pode também ser obtida usando a instrução

```
bayesx_logfile(exempl1_BayesX)
```

8.6 Estudo da convergência: os software CODA e BOA

Como se referiu na secção 6.5, os métodos MCMC visam gerar iterativamente realizações de uma cadeia de Markov homogénea cuja distribuição de equilíbrio é a distribuição *a posteriori* $h(\theta|x)$. O objetivo é obter uma (ou mais) sequência(s) de valores dos parâmetros que possa(m) ser considerada(s) uma amostra representativa da distribuição *a posteriori* conjunta. Para obter, através destes métodos, uma tal amostra é preciso garantir que, a partir de

uma determinada iteração t, a cadeia gerada já se encontra no (ou "próxima" do) estado de equilíbrio. Nesse caso, os estados aí gerados e nas iterações seguintes podem ser considerados como realizações da distribuição *a posteriori* $h(\theta|x)$, naturalmente correlacionadas devido à natureza markoviana da cadeia. Há vários métodos, gráficos e baseados em testes estatísticos, que permitem avaliar a convergência da cadeia ou cadeias geradas para a distribuição estacionária e que foram já referidos na secção 6.5. Os pacotes `CODA` (Plummer *et al*, 2006) e `BOA` (Smith, 2007) implementam estes métodos, permitindo fazer de um modo rápido e eficiente o estudo da convergência.

8.6.1 Testes de diagnóstico de convergência

Tanto o software `CODA` como o `BOA` permitem a utilização dos testes de diagnóstico de convergência, nomeadamente os métodos de Gelman e Rubin (1992), Geweke (1992), Raftery e Lewis (1992), Heidelberg e Welch (1983), os quais vão ser aqui descritos muito brevemente. Uma descrição mais detalhada pode ser encontrada nos artigos referidos, em Cowles e Carlin (1996), ou em Paulino *et al.* (2018).

Método de Gelman e Rubin

Como método para diagnóstico de convergência, Gelman e Rubin sugerem a utilização das componentes da variância de sequências múltiplas da cadeia, simuladas a partir de uma variedade de pontos iniciais dispersos. Para aplicar o método sugerem os seguintes passos:

1. Simulam-se $m \geq 2$ sequências, cada uma de comprimento $2n$, a partir de pontos iniciais simulados de uma distribuição sobredispersa relativamente à distribuição-alvo (distribuição de equilíbrio).

2. Descartam-se as n primeiras iterações de cada sequência.

3. Seja g a quantidade escalar de interesse que se pretende estimar (g é tipicamente uma função do parâmetro θ).

4. Com base nos valores de g calculam-se as componentes da variância W e B, isto é, a variância dentro de cada sequência e a variância entre as sequências, respetivamente.

5. Estima-se a média-alvo de g como uma média amostral de todos os mn valores simulados de g,

6. Estima-se V, a variância-alvo de $g(\theta)$, como uma média ponderada de W e B.

7. Calcula-se o *factor de redução da escala* $\hat{R} = \sqrt{V/W}$.

8. Esta razão decresce para 1 quando $n \to \infty$. Valores de $\hat{R} \approx 1$ são um indício de que cada uma das m sequências de n observações simuladas se aproximam da distribuição-alvo.

Método de Geweke

Seja $\theta^t, t = 1, ..., N$ uma sequência de valores simulados pelo procedimento MCMC e $g(\theta)$ uma função de θ que se pretende estimar. A trajectória $g^1, g^2, \ldots$ construída a partir de $g^t = g(\theta^t)$, define uma série temporal.

O método de Geweke (1992) baseia-se na aplicação de técnicas usuais em séries temporais para averiguar a convergência da sequência simulada. Observa-se a série ao longo de um número N suficientemente longo de iterações e calcula-se a média $g_a = \frac{1}{n_a} \sum g(\theta^t)$ à custa de n_a das primeiras iteradas, bem como a média $g_b = \frac{1}{n_b} \sum g(\theta^t)$ à custa de n_b das últimas iteradas. Se a cadeia é estacionária, então a média da primeira parte da cadeia deve ser semelhante à média da segunda parte da cadeia. Admitindo que n_a/N e n_b/N são fixos e $N \to \infty$ pode mostrar-se que

$$\frac{(g_a - g_b)}{\sqrt{(s_a^2/n_a) + (s_b^2/n_b)}} \quad \to N(0,1),$$

onde s_a^2 e s_b^2 são estimativas independentes das variâncias assintóticas de g_a e g_b, ajustadas em relação à autocorrelação. De acordo com o resultado desta estatística pode averiguar-se se há ou não indicação de convergência.

Método de Raftery e Lewis

Suponha-se que se quer estimar um quantil *a posteriori* q de uma função do parâmetro, com uma certa tolerância r e uma probabilidade s de estar dentro desses limites de tolerância. O método de Raftery e Lewis calcula o número

de iterações N e o número de iterações do período de aquecimento M necessárias para satisfazer as condições especificadas. O resultado do método de diagnóstico tem como componentes, para além de N e M, N_{min} como o número mínimo para uma amostra-piloto e $I = (M+N)/N_{min}$ denominado fator de dependência, interpretado como o incremento proporcional no número de iterações atribuível à dependência serial. Valores elevados deste fator (>5) podem indicar valores iniciais influentes, correlação elevada entre os coeficientes ou uma cadeia com fraca mistura no suporte da distribuição *a posteriori*. Este método deve ser usado com uma ou mais cadeias piloto.

Método de Heidelberg e Welch

Heidelberg e Welch propuseram uma estatística de teste, baseada no teste estatístico de Cramer-von Mises, para testar a hipótese nula de que a cadeia de Markov simulada provém da distribuição estacionária.

O método de diagnóstico, aplicado a cada variável monitorizada, desenvolve-se do seguinte modo:

1. Gera-se uma cadeia de dimensão N e define-se um nível α.

2. Para cada variável monitorizada, calcula-se o valor da estatística de teste usando as N iteradas. De acordo com o resultado do teste toma-se a decisão sobre a rejeição ou não da hipótese nula.

3. Se se rejeitar a hipótese nula, calcula-se de novo a estatística de teste descartando-se 10% das primeiras iteradas. Este procedimento é repetido caso se continue a rejeitar a hipótese nula.

4. Se se continuar a rejeitar a hipótese nula quando o número de iterações usadas no cálculo da estatística de teste atingir os 50% das N iniciais, então o processo iterativo tem de continuar pois a cadeia não atingiu ainda o equilíbrio. Neste caso, CODA dá como resultado a estatística de teste e indica que a cadeia falhou o teste de estacionariedade.

5. Caso contrário, a porção da cadeia responsável pela não rejeição é usada para estimar a média (m) e o erro padrão assintótico (s) da média (calculado usando um método de séries temporais), e sujeita com base nestes valores a um teste adicional rotulado de semiamplitude (*half-width test*) do seguinte modo. Se $1.96 * s < m \times \epsilon$, com ϵ pequeno (CODA

usa por omissão $\alpha = 0.05$ e $\epsilon = 0.1$), então a cadeia passa o teste global. Se $1.96 * s \geq m \times \epsilon$ tal significa que há necessidade de continuar com o processo iterativo.

8.6.2 Os pacotes CODA e BOA

O pacote CODA foi escrito originalmente para o S-PLUS, como parte da tese de doutoramento em Bioestatística de Cowles (1994). Posteriormente foi continuado pela equipa do projeto BUGS (Best *et al.*, 1995), permitindo escrever os valores simulados através do BUGS num formato "coda" com o intuito de serem seguidamente analisados pelo software CODA. O pacote CODA para o R surgiu de uma tentativa de transportar para o ambiente R as funções escritas para o S-PLUS. Dificuldades nessa transposição levaram a uma substancial reescrita dessas funções, e o surgimento do software BOA. O desenvolvimento deste último iniciou-se em 2000 precisamente com a reescrita completa das funções do software CODA para que pudesse ser utilizado em ambiente R. Contudo, atualmente o software CODA já está escrito de modo a ser possível usá-lo em ambiente R (Plummer *et al.*, 2006).

Ambos os pacotes podem ser instalados diretamente do CRAN através das instruções:

```
install.packages("coda",repos="http://cran.us.r-project.org")
install.packages("boa",repos="http://cran.us.r-project.org")
```

Ambos os pacotes CODA e BOA do CRAN têm uma função `codamenu()` e `boa.menu()`, respetivamente, que permitem a utilização do software respetivo em estilo *menu*, para utilizadores casuais que tenham um conhecimento limitado de R. Exemplificando para o CODA:

```
> library(coda)
> codamenu()
CODA startup menu

1: Read BUGS output files
2: Use an mcmc object
3: Quit

Selection: 2
```

```
Enter name of saved object (or type "exit" to quit)
1:A1_2_mcmc #ver posteriormente como foi definido.
Checking effective sample size ...OK
CODA Main Menu

1: Output Analysis
2: Diagnostics
3: List/Change Options
4: Quit
```

Após a leitura dos dados segue uma lista de opções de análise, estatísticas sumárias, representações gráficas e testes de diagnóstico de convergência.

O *menu* do BOA é bastante semelhante. Ver o funcionamento do `boa.menu()` em Smith (2007).

Alternativamente a uma análise através de *menus*, é possível fazer a análise usando comandos do R. De modo a poder construir a interface entre o CODA e o R, no topo de uma infraestrutura baseada em objetos, para que as funções de diagnóstico possam ser utilizadas em modo de linhas de comando de R, em vez de em modo de *menu*, foi criada no R uma nova classe `mcmc`. Esta função aceita como dados de entrada um vetor ou matriz contendo valores simulados resultantes de um procedimento MCMC.

Na última versão dos manuais do CODA e BOA disponíveis em

```
https://cran.r-project.org/web/packages/coda/coda.pdf
https://cran.r-project.org/web/packages/boa/boa.pdf
```

estão descritas as funções para sumariar e representar graficamente os resultados de simulações obtidas com recurso a métodos MCMC, assim como testes de diagnóstico de convergência para a distribuição de equilíbrio das cadeias.

Para utilizar as funções do CODA, tem de ser criado um objeto da classe `mcmc` contendo os valores simulados dos parâmetros a monitorizar. Isso pode ser facilmente executado aplicando a função `as.mcmc()` a uma matriz de dados contendo a informação dos parâmetros a monitorizar como colunas e as iteradas sucessivas como linhas.

Para utilizar as funções do BOA, as cadeias com os parâmetros a monitorizar podem ser dadas através de uma matriz contendo a informação dos parâmetros

como colunas e as iteradas como linhas. A lista com os nomes das linhas e dos parâmetros, tem que fazer parte do argumento `dimnames()` do objeto da classe `matrix`.

8.6.3 Exemplo de aplicação: estudo da convergência com recurso ao CODA e BOA

Vai-se aqui exemplificar como utilizar os programas do `CODA` e `BOA` para estudar a convergência das cadeias anteriormente simuladas usando os pacotes `R2OpenBUGS`, `R2jags`, `RStan` e `R2BayesX`.

A. Usando resultados do R2OpenBUGS

Tal como se viu em 8.2.1, a matriz dos valores simulados dos parâmetros, previamente definidos para monitorização da convergência, obtém-se através da componente $sims.matrix do objeto `A1=Cexemplo1_OpenBUGS.fit` contendo o resultado da função `bugs()`, como se exemplifica a seguir.

```
Cexemplo1_OpenBUGS.fit<- bugs(data=Cexemplo1.data,
inits=list(Inits1,Inits2), parameters.to.save=Cexemplo1.params1,
"Cexemplo1BUGS_semb.txt", n.chains=2, n.iter=40000,
n.burnin=20000, debug=FALSE,save.history=FALSE,DIC=TRUE)

> A1<-Cexemplo1_OpenBUGS.fit$sims.matrix
> dim(A1)
[1] 40000    10
> head(round(A1,4))
     beta0 beta[1] beta[2] beta[3] beta[4]    tau  tau_a
[1,] 15.73   6.349  0.0098   3.843 -0.1046 0.0819 0.0192
[2,] 18.55   2.689  0.0214   4.742 -0.1315 0.0953 0.0195
[3,] 16.41   6.330  0.2284   4.585 -0.0643 0.0664 0.0218
[4,] 14.18   5.653 -0.1744   4.911 -0.1551 0.0793 0.0127
[5,] 19.86   2.291  0.0826   4.259 -0.1209 0.0875 0.0180
[6,] 19.00   1.449 -0.0214   5.277 -0.0964 0.0778 0.0190
     sigma_a sigma deviance
[1,]   7.207 3.495    797.3
[2,]   7.168 3.240    792.8
[3,]   6.779 3.882    800.5
[4,]   8.883 3.552    781.2
```

```
[5,]   7.452 3.381    793.0
[6,]   7.247 3.585    783.9
> class(A1)
[1] "matrix"
```

Vê-se que $A1$ tem 40000 linhas (20000 iteradas para cada uma das 2 cadeias) e 10 colunas com os nomes dos parâmetros monitorizados. Note-se que as iteradas da primeira cadeia estão nas linhas de 1 a 20000 e da segunda cadeia nas linhas de 20001 a 40000.

Assim, para usar o `CODA` com duas cadeias, deve-se definir dois objetos `mcmc`, um para cada cadeia e depois juntá-los num único objeto usando a função `as.mcmc.list`, como se exemplifica seguidamente.

```
> library(coda)
> A1_1chain<-as.mcmc(A1[1:20000,])
> A1_2chain<-as.mcmc(A1[20001:40000,])
> A1_2_mcmc<-as.mcmc.list(list(A1_1chain,A1_2chain))
```

Representações gráficas do traço sobreposto das cadeias e da densidade *a posteriori* (exemplificado na Figura 8.3 para os parâmetros β_1 e β_2), de autocorrelação, da evolução dos quantis e da matriz de correlação entre as componentes da matriz, obtêm-se usando as instruções:

```
plot(A1_mcmc)
plot(A1_2_mcmc[,2:3])    #só para algumas colunas exemplificado na figura
autocorr.plot(A1_2_mcmc) #autocorrelação
cumuplot(A1_2_mcmc)      #evolução dos quantis (0.025,0.5,0.975)
crosscorr.plot(A1_2_mcmc)#imagem da matriz de correlação
```

Os testes de diagnóstico de convergência descritos em 8.6.1 são facilmente executados através das funções respetivas do `CODA`.

1. Método de Gelman e Rubin

```
> gelman.diag(list(A1_1chain,A1_2chain),
confidence = 0.95, transform=FALSE, autoburnin=TRUE,
multivariate=TRUE)
Potential scale reduction factors:

         Point est. Upper C.I.
```

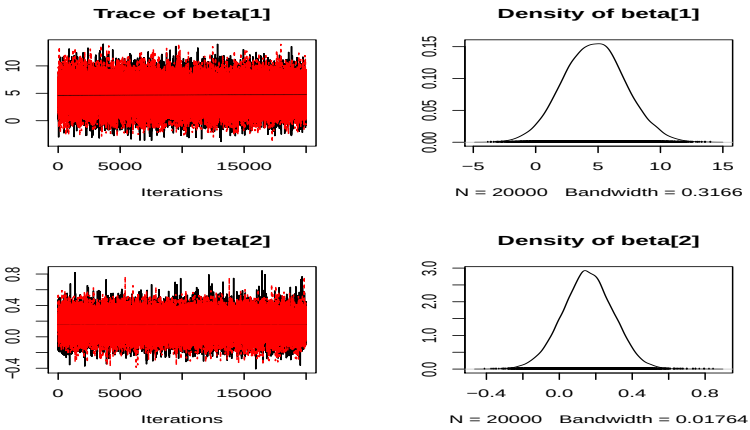

Figura 8.3: Gráfico de traços sobrepostos e densidades *a posteriori* de β_1 e β_2

```
beta0              1          1
beta[1]            1          1
beta[2]            1          1
beta[3]            1          1
beta[4]            1          1
tau                1          1
tau_a              1          1
sigma_a            1          1
sigma              1          1
deviance           1          1

Multivariate psrf
1
```

Como se definiu `autoburnin=TRUE`, só se usou no cálculo dos fatores de redução de escala a segunda metade das séries. Como estes fatores são iguais a 1, tal indica que houve convergência das séries. Uma representação gráfica da evolução do fator de redução de escala de Gelman e Rubin ao longo das iteradas, pode obter-se com a função `gelman.plot()`. Na Figura 8.4 exemplifica-se esta representação para β_1 e β_2.

2. Método de Geweke

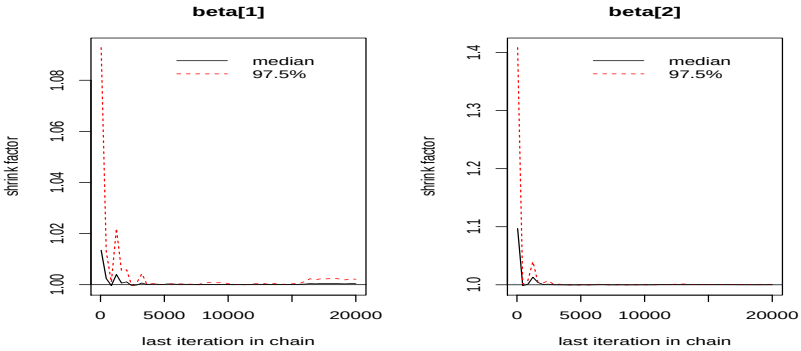

Figura 8.4: Evolução do fator de redução de Gelman e Rubin

```
> geweke.diag(A1_2_mcmc)
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

    beta0   beta[1]   beta[2]   beta[3]   beta[4]       tau
 0.003827 -1.383019 -0.608487 -0.695510 -0.948153  0.047654
    tau_a   sigma_a     sigma  deviance
-0.231638 -0.349071  0.069021 -0.278292


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

   beta0  beta[1]  beta[2]  beta[3]  beta[4]      tau
 -2.2450   2.7206   0.6372  -0.5620  -0.5291   0.3862
   tau_a  sigma_a    sigma deviance
  0.1276  -0.4537  -0.3752   1.4152
```

Os resultados de saída desta função são os *Z-scores* relativos ao teste de igualdade das médias entre a primeira e a última parte das cadeias para cada variável. Dado que, para a primeira cadeia, estes valores estão no intervalo

(-1.96,1.96) não se rejeita a hipótese nula da igualdade das médias para todos os parâmetros monitorizados. Porém, o mesmo não acontece para β_0 e β_1 relativamente à segunda cadeia.

3. Método de Raftery e Lewis

Este método foi desenhado para ser usado numa curta sequência-piloto da cadeia de Markov. Assim vai-se aplicá-lo apenas às 4000 primeiras iteradas de cada cadeia (na realidade as iteradas do período de aquecimento, neste caso 20000, não foram guardadas no objeto A1). Este método exige a definição do quantil que se pretende estimar. Se esse quantil não for definido, o CODA usa por omissão o quantil 0.025. Vai-se aplicar o método para o caso em que se pretende estimar a mediana, ou seja, o quantil 0.5.

```
> raftery.diag(A1_1chain[1:4000,],q=0.5,r=0.01,s=0.95,
converge.eps=0.001)

Quantile (q) = 0.5
Accuracy (r) = +/- 0.01
Probability (s) = 0.95

You need a sample size of at least 9604 with
these values of q, r and s

#usando 10000 iteradas
> raftery.diag(A1_1chain[1:10000,],q=0.5,r=0.01,s=0.95)

Quantile (q) = 0.5
Accuracy (r) = +/- 0.01
Probability (s) = 0.95

          Burn-in  Total Lower bound  Dependence
          (M)      (N)   (Nmin)       factor (I)
 beta0    2        9324  9604         0.971
 beta[1]  2        9268  9604         0.965
 beta[2]  2        9354  9604         0.974
 beta[3]  1        9619  9604         1.000
 beta[4]  2        9099  9604         0.947
 tau      2        9354  9604         0.974
 tau_a    2        9520  9604         0.991
 sigma_a  2        9558  9604         0.995
```

```
 sigma    2        9384  9604         0.977
 deviance 2        9272  9604         0.965
```

```
> raftery.diag(A1_2chain[1:10000,],q=0.5,r=0.01,s=0.95)

Quantile (q) = 0.5
Accuracy (r) = +/- 0.01
Probability (s) = 0.95

          Burn-in  Total Lower bound  Dependence
          (M)      (N)   (Nmin)       factor (I)
 beta0    2        9794  9604         1.020
 beta[1]  2        9771  9604         1.020
 beta[2]  2        9459  9604         0.985
 beta[3]  1        9588  9604         0.998
 beta[4]  2        9302  9604         0.969
 tau      2        9736  9604         1.010
 tau_a    2        9406  9604         0.979
 sigma_a  2        9399  9604         0.979
 sigma    2        9751  9604         1.020
 deviance 2        9276  9604         0.966
```

O fator de dependência para ambas as cadeias é próximo de 1, não indicando problemas. De acordo com os resultados, as 10000 iteradas seriam suficientes para estimar a mediana com uma tolerância $r = 0.01$ e uma probabilidade $s = 0.95$ de estar dentro desses limites de tolerância.

4. Método de Heidelberg e Welch

Para aplicação deste método deve-se fixar o valor de ϵ e α. Como se disse anteriormente o CODA usa, por omissão, os valores 0.1 e 0.05, respetivamente. Esses valores podem ser alterados nos argumentos da função `heidel.diag()`. Vai-se usar um valor de $\epsilon = 0.01$ apenas para exemplificação.

```
> heidel.diag(A1_2_mcmc, eps=0.01, pvalue=0.05)
[[1]]

         Stationarity start     p-value
         test         iteration
beta0    passed       1         0.503
beta[1]  passed       1         0.052
beta[2]  passed       1         0.592
```

```
beta[3]  passed       1         0.822
beta[4]  passed       1         0.504
tau      passed       1         0.402
tau_a    passed       1         0.999
sigma_a  passed       1         0.936
sigma    passed       1         0.435
deviance passed       1         0.503

         Halfwidth Mean     Halfwidth
         test
beta0    passed     17.2739 2.47e-02
beta[1]  passed      4.7013 3.46e-02
beta[2]  failed      0.1565 1.93e-03
beta[3]  passed      4.2236 1.29e-02
beta[4]  passed     -0.1047 4.93e-04
tau      passed      0.0774 1.52e-04
tau_a    passed      0.0163 5.29e-05
sigma_a  passed      7.9973 1.30e-02
sigma    passed      3.6216 3.60e-03
deviance passed    794.9228 1.72e-01

[[2]]

         Stationarity start     p-value
         test         iteration
beta0    passed       2001      0.2585
beta[1]  passed          1      0.0766
beta[2]  passed          1      0.8299
beta[3]  passed          1      0.1795
beta[4]  passed          1      0.8124
tau      passed          1      0.9457
tau_a    passed          1      0.8847
sigma_a  passed          1      0.9781
sigma    passed          1      0.9562
deviance passed          1      0.5130

         Halfwidth Mean     Halfwidth
         test
beta0    passed     17.2694 0.025928
beta[1]  passed      4.6858 0.034335
```

```
beta[2]  failed      0.1561 0.001879
beta[3]  passed      4.2275 0.012846
beta[4]  passed     -0.1046 0.000507
tau      passed      0.0773 0.000154
tau_a    passed      0.0162 0.000052
sigma_a  passed      8.0069 0.013041
sigma    passed      3.6251 0.003658
deviance passed    795.0646 0.175413
```

Vê-se que com este valor de $\epsilon = 0.01$ o parâmetro β_2 passava o teste de estacionariedade para ambas as cadeias, mas falhava o teste de semiamplitude, revelando a necessidade de continuar com o processo iterativo para atingir a precisão exigida. Se se aumentar ϵ para 0.05 já todos os parâmetros passam os dois testes.

5. Intervalos HPD

Finalmente relembre-se que com o CODA podem-se obter intervalos HPD para todos os parâmetros monitorizados usando a função `HPDinterval()`, exemplificando-se aqui a obtenção de intervalos HPD de 95% de credibilidade, para cada uma das cadeias.

```
> HPDinterval(A1_2_mcmc, prob = 0.95)
[[1]]
              lower     upper
beta0     13.860000  20.80000
beta[1]    0.036880   9.84600
beta[2]   -0.128400   0.42490
beta[3]    2.487000   6.08700
beta[4]   -0.178200  -0.03300
tau        0.056580   0.10000
tau_a      0.009279   0.02371
sigma_a    6.246000   9.88700
sigma      3.122000   4.14500
deviance 771.400000 820.30000
attr(,"Probability")
[1] 0.95

[[2]]
              lower     upper
beta0     13.960000  20.87000
```

```
beta[1]    0.002304   9.64300
beta[2]   -0.114500   0.42770
beta[3]    2.402000   6.02500
beta[4]   -0.178100  -0.03235
tau        0.055750   0.09912
tau_a      0.009449   0.02374
sigma_a    6.223000   9.83100
sigma      3.122000   4.15300
deviance 771.100000 819.80000
attr(,"Probability")
[1] 0.95
```

Vai-se agora exemplificar, muito rapidamente, como os mesmos métodos podem ser aplicados usando o programa BOA.

```
A1<-Cexemplo1_OpenBUGS.fit$sims.matrix #resultado na forma de matriz
nomes<-list(c(1:20000,1:20000),c("beta0","beta[1]","beta[2]",
"beta[3]","beta[4]","tau","tau_a","sigma_a","sigma","deviance"))
dimnames(A1)<-nomes
A1_1<-A1[1:20000,]#define a primeira cadeia
A1_2<-A1[20001:40000,]#define a segunda cadeia
#-----------------------------------------------#
        #autocorrelação#
#-----------------------------------------------#
boa.acf(A1_1,lags=1)
boa.acf(A1_2,lags=1)
#-----------------------------------------------#
        #método Geweke #
#-----------------------------------------------#
boa.geweke(A1_1, p.first=0.1, p.last=0.5)
boa.geweke(A1_2, p.first=0.1, p.last=0.5)
#-----------------------------------------------#
     #Método de Heidelberg e Welch #
#-----------------------------------------------#
boa.handw(A1_1, error=0.05, alpha=0.05)
boa.handw(A1_2, error=0.05, alpha=0.05)
#-----------------------------------------------#
            #Intervalos HPD #
#-----------------------------------------------#
#a função boa.hpd() dá o intervalo HPD
#para um único parâmetro.
```

```
#para definir intervalos simultâneamente
#para todos os parâmetros monitorizados
#podemos proceder como segue
hpd_boa<-function(x) boa.hpd(x,0.05)
apply(A1_1,2,hpd_boa)
apply(A1_2,2,hpd_boa)
```

B. Usando o R2jags

1. O resultado do modelo usando a função `jags()` pode ser convertido num objeto MCMC, o qual pode ser gerado para análise através do comando:

```
exemplo1_JAGS.fit2.mcmc <- as.mcmc(exemplo1_JAGS.fit2 )
```

2. Como anteriormente, com este objeto MCMC pode usar-se uma variedade de comandos para diagnóstico usando o CODA:

```
library(coda)
plot(exemplo1_JAGS.fit2.mcmc)
autocorr.plot(exemplo1_JAGS.fit2.mcmc)
gelman.plot(exemplo1_JAGS.fit2.mcmc)
gelman.diag(exemplo1_JAGS.fit2.mcmc)
geweke.diag(exemplo1_JAGS.fit2.mcmc)
raftery.diag(exemplo1_JAGS.fit2.mcmc)
heidel.diag(exemplo1_JAGS.fit2.mcmc)
```

Note-se que o objeto `exemplo1_JAGS.fit2` já continha as duas cadeias, o mesmo acontecendo ao objeto `mcmc` definido por `exemplo1_JAGS.fit2.mcmc`, não sendo pois necessário discriminar as duas cadeias, como acontecia com o resultado do R2OpenBUGS.

3. A função `jags()` também retorna os valores simulados dos parâmetros monitorizados no formato matricial em:

```
exemplo1_JAGS.fit2$BUGSoutput$sims.matrix
```

Como tal, para usar o programa BOA, procede-se exatamente como indicado para a situação em que as cadeias foram geradas através do R2OpenBUGS.

C. Usando o RStan

Como se viu no ponto 5 da subsecção 8.4.1, os valores simulados dos parâmetros gerados usando a função `stan()` podem ser obtidos aplicando a função `extract()` ao objeto criado por aquela função.

```
> samples_stan_array<-extract(exemplo1.fit_stan,
+ pars=c("beta0", "beta", "sigma", "sigma_a", "tau", "tau_a"),
+ permuted = FALSE, inc_warmup = FALSE, include = TRUE)
> class(samples_stan_array)
[1] "array"
> dim(samples_stan_array)
[1] 20000     2     9 #20000 cada cadeia, 2 cadeias, 9 parâmetros
```

Para usar o `CODA` ou o `BOA` tem-se de começar por definir matrizes contendo cada uma das cadeias.

```
samples_coda_1<-as.matrix(samples_stan_array[1:20000,1,1:9]))
samples_coda_2<-as.matrix(samples_stan_array[1:20000,2,1:9]))
```

Para usar o `CODA` elas devem ser transformadas em objetos da classe `mcmc`.

```
samples_coda_1<-mcmc(samples_coda_1)
samples_coda_2<-mcmc(samples_coda_2)
gelman.diag(list(samples_coda_1,samples_coda_2))
geweke.diag(samples_coda_1)
geweke.diag(samples_coda_2)
raftery.diag(samples_coda_1)
raftery.diag(samples_coda_2)
heidel.diag(samples_coda_1)
heidel.diag(samples_coda_2)
```

Para usar o `BOA` tem-se de definir previamente em `dimnames` os nomes das linhas e colunas.

```
samples_coda_1<-as.matrix(samples_stan_array[1:20000,1,1:9])
samples_coda_2<-as.matrix(samples_stan_array[1:20000,2,1:9])
dimnames(samples_coda_1)<-list(1:20000, c("beta0", "beta[1]",
"beta[2]", "beta[3]", "beta[4]", "sigma", "sigma_a", "tau", "tau_a"))
dimnames(samples_coda_2)<-list(1:20000, c("beta0", "beta[1]",
"beta[2]", "beta[3]", "beta[4]", "sigma", "sigma_a", "tau", "tau_a"))
```

Exemplificando para o método de Geweke

```
> boa.geweke(samples_coda_1,p.first=.1,p.last=0.5)
           Z-Score    p-value
beta0   -0.1895212 0.84968432
beta[1] -1.1536020 0.24866338
beta[2] -0.3998341 0.68927871
beta[3] -0.3581599 0.72022368
beta[4] -1.3735690 0.16957554
sigma   -0.7696775 0.44149123
sigma_a -1.7314080 0.08337903
tau      0.6671540 0.50467380
tau_a    1.7048218 0.08822767
> boa.geweke(samples_coda_2,p.first=.1,p.last=0.5)
            Z-Score   p-value
beta0   -0.51871293 0.6039609
beta[1]  0.15164978 0.8794632
beta[2]  1.35185008 0.1764233
beta[3] -0.57649303 0.5642820
beta[4]  0.61505637 0.5385175
sigma   -0.93391998 0.3503452
sigma_a -0.03298591 0.9736858
tau      1.23723600 0.2159995
tau_a    0.02936042 0.9765771
```

D. Usando o R2BayesX

Recorde-se o ponto 4 da subsecção 8.5.1. Viu-se que a função `samples()` aplicada ao objeto resultante da função `bayesx()` retorna os valores simulados da distribuição *a posteriori* dos parâmetros indicados no argumento `term`. O argumento `CODA` controla o tipo de classe do objeto resultante.

Assim com a instrução seguinte obtém-se uma lista `mcmc` que pode ser usada para fazer o estudo de diagnóstico de convergência através do `CODA`:

```
> AA_coda<-samples(exemplo1_BayesX,
+ term=c("linear-samples","var-samples","sd(ID)"),coda=TRUE)
> class(AA_coda)
[1] "mcmc.list"
> names(AA_coda)
```

```
[1] "Chain_1" "Chain_2"

#exemplificando

> gelman.diag(AA_coda)
Potential scale reduction factors:

          Point est. Upper C.I.
Intercept          1          1
z2                 1          1
z3                 1          1
z1                 1          1
z                  1          1
                   1          1

Multivariate psrf
1
```

Por outro lado, para utilizar o programa BOA para o estudo de convergência procede-se como segue:

```
> AA_boa<-as.matrix(AA_data)
> AA_boa_1<-AA_boa[,1:6]
> AA_boa_2<-AA_boa[,7:12]

#exemplificando
> library(boa)
> boa.geweke(AA_boa_1,p.first=0.1,p.last=0.5)
                           Z-Score   p-value
Chain_1.Param.Intercept -0.2281478 0.8195313
Chain_1.Param.z2         1.2278951 0.2194863
Chain_1.Param.z3        -0.2358216 0.8135711
Chain_1.Param.z1         1.1215734 0.2620438
Chain_1.Param.z          0.8195813 0.4124548
Chain_1.Var              0.6110576 0.5411614
> boa.geweke(AA_boa_1,p.first=0.1,p.last=0.5)
                           Z-Score   p-value
Chain_1.Param.Intercept -0.2281478 0.8195313
Chain_1.Param.z2         1.2278951 0.2194863
Chain_1.Param.z3        -0.2358216 0.8135711
Chain_1.Param.z1         1.1215734 0.2620438
```

```
Chain_1.Param.z          0.8195813 0.4124548
Chain_1.Var              0.6110576 0.5411614
```

8.7 R-INLA e exemplo de aplicação

Na secção 7.3 descreveu-se o método INLA para analisar modelos bayesianos hierárquicos sem recurso a métodos de simulação. Sendo um método de aproximação, com esta metodologia não há necessidade de haver preocupações com os problemas de convergência inerentes aos métodos de simulação MCMC, avaliados na secção anterior. No entanto, isso não significa que as aproximações obtidas para as distribuições *a posteriori* dos parâmetros do modelo sejam sempre boas, havendo pois necessidade de estudar a qualidade da aproximação. Rue *et al.* (2009) propuseram duas estratégias para avaliar o erro de aproximação da distribuição *a posteriori*: uma baseada no cálculo do número efetivo de parâmetros e outra no uso do critério de divergência de Kullback-Leibler. Podem encontrar-se detalhes destas estratégias nas secções 4.1 e 4.2 do referido artigo. Estas estratégias estão implementadas no `R-INLA`. Posteriormente vai-se ver como avaliá-las.

Como se referiu no Capítulo 7, a metodologia `INLA` é adequada para fazer inferência bayesiana em modelos gaussianos latentes, uma classe de modelos bastante flexível que engloba desde modelos aditivos mistos linear generalizados a processos de Cox log-gaussianos e modelos espaço-temporais. Combinando com a abordagem de equações diferenciais parciais estocásticas (SPDE) - *vide* Lindgren *et al.*, 2011 -, pode modelar-se todo o tipo de dados geograficamente referenciados, incluindo dados de áreas, dados georeferenciados e dados de processos pontuais espaciais (Lindgren e Rue, 2015)

O software `R-INLA` é um pacote do `R` desenvolvido para implementar inferência bayesiana aproximada usando a metodologia `INLA`. Este pacote aparece no seguimento do programa autónomo INLA, escrito na linguagem `C` e construído sobre a programoteca `GMRFLib` de `C` - veja-se

http://www.math.ntnu.no/ hrue/GMRFLib/doc/html/,

destinada a uma rápida e exata simulação de campos aleatórios gaussianos.

O `R-INLA` está disponível para os sistemas operativos Linux, Mac e Windows. No sítio `www.r-inla.org`, para além de instruções de como instalar o `R-INLA`, encontram-se códigos, exemplos, artigos e relatórios onde se discute

a teoria e aplicações do INLA, e muito outro material de grande interesse, nomeadamente um fórum de discussão e uma coleção de respostas às perguntas mais frequentes.

Para instalar o `R-INLA` diretamente da consola do **R** escreve-se a instrução:

```
install.packages("INLA",
 repos="http://www.math.ntnu.no/inla/R/stable")
```

Como para qualquer outra programoteca do `R`, para carregar o `R-INLA` em cada sessão de trabalho escreve-se:

```
library(INLA)
```

Como existem constantes atualizações do `R-INLA`, deve usar-se a função

```
inla.upgrade(testing=FALSE)
```

para obter a última versão mais estável do pacote.

Há uma grande variedade de distribuições de probabilidade que podem ser usadas para a variável resposta com recurso ao `R-INLA`. Pode obter-se essa lista através da instrução

```
> names(inla.models()$likelihood)
```

Em http://www.r-inla.org/models/likelihoods encontra-se uma descrição completa dessas distribuições com exemplos de aplicação.

Do mesmo modo, para obter informação sobre as distribuições *a priori* para os parâmetros e para os efeitos estruturados e não estruturados, pode consultar-se

```
http://www.r-inla.org/models/priors
http://www.r-inla.org/models/latent-models
```

As listas com as correspondentes distribuições podem ainda ser obtidas através das instruções

```
>  names(inla.models()$prior)
>  names(inla.models()$latent)
```

Para melhor compreender como funciona o `R-INLA` passa-se ao estudo do exemplo que tem vindo a ser usado.

8.7.1 Exemplo de aplicação

1. Tal como para o R2BayesX, depois de se term especificado o modelo bayesiano, tal como foi feito na secção 8.1, o modelo é traduzido para R através da construção do objeto que contém a fórmula

```
 > INLA_formula <- X ~ z1 + z2 + z3+ z+
+           f(ID, model="iid",
hyper=list(prec=list(prior="loggamma",param=c(1,0.005))))
> class(INLA_formula)
[1] "formula"
```

Como anteriormente, as variáveis correspondentes aos efeitos fixos e que aparecem linearmente no modelo, foram previamente centradas. Através da função f() que aparece na definição da fórmula, são estabelecidos os efeitos estruturados (os vários tipos estão definidos em

http://www.r-inla.org/models/latent-models).

No caso em mão tem-se apenas os efeitos aleatórios relativos ao indivíduo (variável ID) e concretizados no modelo através de "a". O modelo "iid" corresponde a uma distribuição normal com valor médio zero e precisão τ_a. A distribuição *a priori* especificada no argumento hyper é para $log(\tau_a)$. Sendo uma log-gama, corresponde a uma distribuição gama para a precisão τ_a .

2. Seguidamente chama-se a funcão inla() para correr o algoritmo INLA e obter os resultados necessários para proceder à inferência bayesiana, como se segue:

```
> ?inla
> resultado_INLA <- inla(INLA_formula,family="normal",
+ control.predictor=list(compute=TRUE),
+ control.compute =list(waic=TRUE,dic=TRUE,cpo=TRUE),
+ data = Cexemplo1
+ control.family=list(hyper=list
+ (prec=list(prior="loggamma",param=c(1,0.005)))))
```

A primeira linha do código acima permite conhecer todos os argumentos da função inla(), dos quais apenas são obrigatórios especificar o objeto que contém a fórmula, neste caso, INLA_formula e o objeto contendo os dados, neste caso data = Cexemplo1. Não especificando os outros argumentos, serão considerados os definidos por omissão pelo R-INLA.

Ao escrever `control.predictor=list(compute=TRUE)` está-se a pedir que sejam calculadas as distribuições marginais do preditor linear. Há outros argumentos desta função, cujo conhecimento se obtém através da instrução

```
?control.predictor
```

Em `control.family()` declaram-se as distribuições *a priori* para os parâmetros da família de distribuições de amostragem. Neste caso declarou-se a distribuição *a priori* para a precisão τ. Veja-se em `?control.family` como se deve fazer essa declaração para parâmetros de certas distribuições.

Para que sejam calculados os critérios WAIC e DIC e também as ordenadas preditivas condicionais (CPO), deve declarar-se

```
control.compute =list(waic=TRUE,dic=TRUE,cpo=TRUE)
```

3. A função `inla()` devolve um objeto da classe `inla`, aqui designado por `resultado_INLA`. Esta objeto é uma lista contendo muitos outros objetos que podem ser explorados usando a instrução `names(resultado_INLA)`. Resultados sumários do procedimento INLA são obtidos através da instrução:

```
> summary(resultado_INLA)
...
Time used:
 Pre-processing    Running inla Post-processing       Total
         0.1719          0.4375          0.0975      0.7069

Fixed effects:
               mean sd 0.025quant 0.5quant 0.975quant mode   kld
(Intercept) 17.250 1.737    13.820  17.251    20.669 17.253   0
z1           4.772 2.448    -0.050   4.770     9.599  4.766   0
z2           0.154 0.138    -0.118   0.154     0.427  0.153   0
z3           4.169 0.908     2.394   4.164     5.972  4.154   0
z           -0.106 0.036    -0.176  -0.106    -0.035 -0.107   0

Random effects:
Name      Model
 ID   IID model

Model hyperparameters:
                                          mean sd 0.025quant 0.5quant
```

```
Precision for the Gaussian observations 0.0789 0.0113  0.0587 0.0782
Precision for ID                        0.0170 0.0039  0.0106 0.0167
                                        0.975quant   mode
Precision for the Gaussian observations     0.1027 0.0771
Precision for ID                            0.0256 0.0160

Expected number of effective parameters(std dev): 46.38(0.8267)
Number of equivalent replicates : 3.17

Deviance Information Criterion (DIC) ...: 841.57
Effective number of parameters .........: 47.65

Watanabe-Akaike information criterion (WAIC) ...: 844.77
Effective number of parameters .................: 41.31

Marginal log-Likelihood:  -497.49
Posterior marginals for linear predictor and fitted values computed
```

Compare-se estes resultados com os obtidos com os métodos de simulação. Em particular, compare-se os valores do WAIC com os obtidos com o `RStan`.

4. Note que nos resultados também é reportado, para cada configuração dos hiperparâmetros, uma estimativa para o número efetivo de parâmetros. Esta estimativa corresponde basicamente ao número esperado de parâmetros independentes do modelo. No nosso caso temos 7+49=56 parâmetros, mas como os efeitos aleatórios são correlacionados, o número esperado de parâmetros independentes é inferior, ≈ 47, como se observa. Como se referiu, esta é uma das estratégias sugeridas por Rue *et al.* (2009) para avaliar exatidão da aproximação. Nomeadamente, se o número efetivo de parâmetros é pequeno comparado com a dimensão da amostra, então espera-se que a aproximação seja boa. Neste caso a razão entre a dimensão da amostra (147) e o número efetivo de parâmetros (46.38) é cerca de 3.17, o que sugere uma razoável qualidade da aproximação. Esta razão é de facto o número de "réplicas equivalentes" o qual corresponde ao número de observações por cada número esperado de parâmetros efetivos.

Outra quantidade que é registada é o valor da medida de discrepância de Kullback-Liebler (na coluna kld). Este valor descreve a diferença entre a aproximação gaussiana e a aproximação simplificada de Laplace (relembre-se o que foi dito no Capítulo 7 sobre as várias estratégias usadas pelo INLA para

obter as aproximações) para as distribuições marginais *a posteriori*. Valores pequenos indicam que a distribuição *a posteriori* é bem aproximada por uma distribuição Normal.

5. A estratégia de aproximação que está implementada, por omissão, através da função `inla()` é a abordagem simplificada de Laplace. Outras estratégias de aproximação e de integração podem ser definidas usando o argumento `control.inla` da função `inla()`. Por exemplo, se se quisesse que fosse implementada a abordagem de Laplace completa, o que é aconselhável fazer para aumentar a exatidão da estimação das caudas das distribuições marginais, acrescentar-se-ia à função `inla()` o argumento

```
control.inla=list(strategy="laplace",npoints=21)
```

6. Para além dos resultados sumários apresentados anteriormente, o `R-INLA` também permite a determinação de dois tipos de medidas de qualidade de ajustamento, nomeadamente as ordenadas preditivas condicionais $p(y_i|y_{-i})$ (CPO) e as probabilidades de transformação uniformizante $P(Y_i^{nova} \leq y_i|y_{-i})$ (PIT de *probability integral transforms*). Para tal basta acrescentar, tal como já se disse, `cpo=TRUE` na lista do argumento `control.compute` da função `inla`. Estes valores fazem parte dos resultados devolvidos pela função `inla`. A lista de todos esses resultados pode obter-se através da instrução `names(resultados_INLA)`. Dos 51 resultados possíveis, ilustra-se apenas alguns:

```
> names(resultado_INLA)
 [1] "names.fixed"
 [2] "summary.fixed"
 [3] "marginals.fixed"
 [4] "summary.lincomb"
 [5] "marginals.lincomb"
 [6] "size.lincomb"
 [7] "summary.lincomb.derived"
 [8] "marginals.lincomb.derived"
 [9] "size.lincomb.derived"
[10] "mlik"
[11] "cpo"
[12] "po"
[13] "waic"
...
[18] "summary.linear.predictor"
[19] "marginals.linear.predictor"
```

```
[20] "summary.fitted.values"
[21] "marginals.fitted.values"
...
[27] "offset.linear.predictor"
...
[51] "model.matrix"
```

Assim pode-se colocar num objeto com o nome, por exemplo, `CPO_PIT` os valores dessas duas medidas para fazer posteriormente um estudo adequado da qualidade de ajustamento, representações gráficas, etc.

```
> CPO_PIT<-resultado_INLA$cpo
> names(CPO_PIT)
[1] "cpo"     "pit"     "failure"
> class(CPO_PIT)
[1] "list"
> summary(CPO_PIT$cpo)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0003652 0.0486900 0.0741300 0.0673100 0.0905200 0.0924000
> summary(CPO_PIT$pit)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0004578 0.2769000 0.5046000 0.5013000 0.7493000 0.9961000
> summary(CPO_PIT$failure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0       0       0       0       0
```

Valores extremos de CPO indicam observações "surpreendentes" e valores extremos de PIT indicam observações discordantes (*outliers*). Um histograma destas probabilidades PIT com um aspeto que não seja consonante com o de uma distribuição uniforme, revela um modelo não adequado. De acordo om a informação contida em `CPO_PIT$failure`, nenhuma das observações é considerada surpresa ou discordante.

7. Para obter representações gráficas pode-se utilizar a função `plot()`. Esta função tem vários argumentos com valor lógico. Por omissão esse valor é "TRUE". Assim, se se escrever

```
plot(resultado_INLA)
#ou para ser em janelas individuais
plot(resultado_INLA,single = TRUE)
```

obtém-se os gráficos das densidades *a posteriori* dos efeitos fixos e aleatórios, das densidades *a posteriori* das precisões, da série das médias e quantis de probabilidade 0.025 e 0.975 *a posteriori* dos efeitos aleatórios e do preditor linear, dos valores CPO e PIT e correspondentes histogramas. Se se quiser executar os gráficos um a um basta escrever "FALSE" no argumento lógico dos elementos cujo gráfico não se quer executar. Por exemplo, para obter apenas das densidades *a posteriori* das precisões escreve-se:

```
plot(resultado_INLA,
plot.fixed.effects = FALSE,
plot.lincomb = FALSE,
plot.random.effects = FALSE,
plot.hyperparameters = TRUE,
plot.predictor = FALSE,
plot.q = FALSE,
plot.cpo = FALSE)
```

obtendo-se a Figura 8.5.

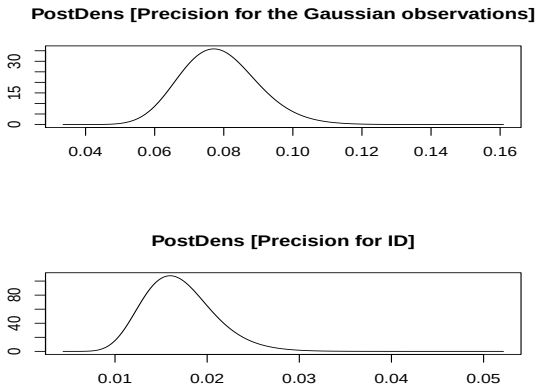

Figura 8.5: Densidade a posteriori de τ e τ_a

8. O R-INLA trabalha com precisões (inverso das variâncias). No entanto, em geral, tem mais interesse fazer inferências sobre o desvio padrão. Isso é possível fazendo uso de um conjunto de funções incluídas no pacote INLA que permitem o cálculo de quantis, percentis, valores esperados de funções dos parâmetros originais, e inclusivamente obter amostras das distribuições

a posteriori marginais. Assim, para obter a média *a posteriori* dos desvios padrões $\sigma = 1/\sqrt{\tau}$ e de $\sigma_a = 1/\sqrt{\tau_a}$, procede-se como se segue:

```
> names(resultado_INLA$marginals.hyperpar)
[1] "Precision for the Gaussian observations"
[2] "Precision for ID"
> tau<-resultado_INLA$marginals.hyperpar$
+"Precision for the Gaussian observations"
> sigma<-inla.emarginal(function(x) 1/sqrt(x), tau)
> sigma
[1] 3.588387
> tau_a<-resultado_INLA$marginals.hyperpar$"Precision for ID"
> sigma_a<-inla.emarginal(function(x) 1/sqrt(x), tau_a)
> sigma_a
[1] 7.813781
```

Alternativamente, para obter a distribuição *a posteriori* dos desvios padrões dos efeitos aleatórios estruturados pode também usar-se a instrução:

```
> sigmas<-inla.contrib.sd(resultado_INLA,nsamples=1000)
> names(sigmas)
[1] "samples" "hyper"
> sigmas$hyper
                                    mean  sd     2.5%   97.5%
sd for the Gaussian observations 3.59079 0.25221 3.1267 4.1120
sd for ID                        7.79427 0.88068 6.2625 9.6867
> head(sigmas$samples)
     sd for the Gaussian observations sd for ID
[1,]                         3.407485  8.287859
[2,]                         3.775560  6.945835
[3,]                         3.912179  9.931287
[4,]                         3.282005 10.068471
[5,]                         3.736729  7.386682
[6,]                         3.808289  9.027061
```

O objeto sigmas acima criado contém em "samples" um vetor de simulações da distribuição *a posteriori* dos desvios padrões.

9. Para obter amostras aleatórias de distribuições *a posteriori* marginais usa-se a função `inla.rmarginal()` do seguinte modo (como exemplificado para a distribuição *a posteriori* de β_3:

```
> names(resultado_INLA$marginals.fixed)
[1] "(Intercept)" "z1"        "z2"          "z3"         "z"
> dens_z3<-resultado_INLA$marginals.fixed$z3
> amostra_z3<-inla.rmarginal(1000,dens_z3)
> summary(amostra_z3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5421  3.5580  4.1580  4.1630  4.7880  7.1570
```

Informação sobre as funções para operar sobre marginais encontra-se usando a instrução `?inla.marginal`.

10. Uma dessas funções, `inla.hpdmarginal()`, permite obter intervalos de credibilidade HPD para os parâmetros do modelo. Para obter esses intervalos para os parâmetros do preditor linear correspondentes aos efeitos fixos, pode proceder-se individualmente, ou em grupo, como se exemplifica aqui para um intervalo de 95%:

```
> HPD<-NULL
> for(i in 1:5){
+ HPD[[i]]<-inla.hpdmarginal
+ (0.95, resultado_INLA$marginals.fixed[[i]])}
> HPD
[[1]]
                low     high
level:0.95 13.82332 20.66469

[[2]]
                   low    high
level:0.95 -0.05260815 9.58653

[[3]]
                  low      high
level:0.95 -0.1184688 0.4263571

[[4]]
                low     high
level:0.95 2.384431 5.958083

[[5]]
                  low        high
level:0.95 -0.1759522 -0.03584334
```

e para os hiperparâmetros do modelo

```
> names(resultado_INLA$marginals.hyper)
[1] "Precision for the Gaussian observations"
[2] "Precision for ID"
> HPDhyper<-NULL
> for(i in 1:2){
+ HPDhyper[[i]]<-inla.hpdmarginal
+ (0.95, resultado_INLA$marginals.hyper[[i]])}
> HPDhyper
[[1]]
                 low      high
level:0.95 0.0574843 0.1011766

[[2]]
                   low       high
level:0.95 0.009948865 0.02469588
```

Como era de esperar os intervalos HPD para os parâmetros dos efeitos fixos praticamente coincidem com os intervalos de caudas iguais obtidos no sumário dos resultados. O mesmo não acontece com os parâmetros de precisão.

Bibliografia

[1] Adler, D., Kneib, T., Lang, S., Umlauf, N. e Zeileis, A. (2013). BayesXsrc: R Package Distribution of the `BayesX` C++ Sources. R package version 2.1-2. https://cran.r-project.org/web/packages/BayesXsrc/

[2] Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Em *Second International Symposium on Information Theory*(B.N. Petrov and F. Csaki, eds.), 267-281. Akademiai Kiado, Budapest.

[3] Amaral Turkman, M.A. (1980). *Applications of Predictive Distributions*. PhD. Thesis, University of Sheffield.

[4] Basu, D. e Pereira, C.A.B. (1982). On the Bayesian analysis of categorical data: the problem of nonresponse. *J. Statist. Plann. Infer.*, **6**, 345–362.

[5] Belitz, C., Brezger, A., Kneib, T., Lang, S. e Umlauf, N. (2013). `BayesX`: Software for Bayesian Inference in Structured Additive Regression Models. Version 2.1. www.bayesx.org/

[6] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Inference*. Springer, Berlim.

[7] Bernardo, J.M. e Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.

[8] Best, N.G., Cowles, M.K. e Vines, S.K. (1996). *CODA: Convergence Diagnostics and Output Analysis Software for Gibbs Sampling Output*. Versão 0.4. MRC Biostatistics Unit, Cambridge, U.K.

[9] Blangiardo, M., Cameletti, M., Baio, G. e Rue, H. (2013) Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*,7, 39-55.

[10] Blangiardo, M. e Cameletti, M. (2015) *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley.

[11] Box, G.E.P. (1980). Sampling and Bayes inference in scientific modelling and robustness. *J. Royal Statist. Soc. A*, **143**, 383-430.

[12] Burnham, K.P. e Anderson, D.R. (2002). *Model Selection and Multimodel Inference: a practical information-theoretic approach*. 2nd ed., Springer, New York.

[13] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P. e Riddell, A. (2015) Stan: A Probabilistic Programming Language. To be published in *Journal of Statistical Software*.

[14] Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.*, **89**, 818–824.

[15] Chen, M.-H. e Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *J. Comput. and Graphical Stat.*, **8**, 69–92.

[16] Chen, M.-H., Shao, Q.-M. e Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

[17] Cowles, M.K. (1994). *Practical issues in Gibbs sampler implementation with application to Bayesian hierarchical modelling of clinical trial data*. PhD thesis, Division of Biostatistics, University of Minnesota.

[18] Cowles, M. e Carlin, B. (1996). Markov Chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.*, **91**, 883-904.

[19] Damien, P., Wakefiel, J. e Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. Royal Statist. Soc. B*, **61(2)**, 331–344.

[20] Dawid, A.P. (1985). The impossibility of inductive inference. (Invited discussion of 'Self-calibrating priors do not exist', by D. Oakes.) *J. Amer. Statist. Assoc.*, **80**, 340–341.

[21] Devroye, L. (1986). *Non-uniform Random Variate Generator*. Springer-Verlag, New York.

[22] Fahrmeir, L. e Tutz, G. (2001) *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer Verlag, Berlin.

[23] Gelfand, A.E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson e D.J. Spiegelhalter, eds.), 145-161. Chapman and Hall, London.

[24] Gelfand, A. E. e Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations, *J. Royal Statist. Soc. B*, **56**, 501-514.

[25] Gelfand, A.E. e Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

[26] Gelfand, A.E., Smith, A.F.M. e Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.*, **87**, 523–531.

[27] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. e Rubin, D.B. (2014b). *Bayesian Data Analysis*. 3rd ed., Chapman and Hall/CRC, Boca Raton, FL.

[28] Gelman, A., Hwang, J. e Vehtari, A. (2014). Understanding predictive information criterion for Bayesian models. *Statistics and Computing*, **24**, 997–1016.

[29] Gelman, A. e Meng, X.L. (1996). Model checking and model improvement. In *Markov Chain Monte Carlo in Practice* (W.R. Gilks, S. Richardson e D.J. Spiegelhalter, eds.), 189-202. Chapman and Hall, London.

[30] Gelman, A. e Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–72.

[31] Geman, S. e Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

[32] Gentle, J.E. (2004). *Random Number Generation and Monte Carlo Methods*. 2nd ed., Springer, New York.

[33] Genz, A. e Kass, R.E. (1997). Subregion adaptative integration of functions having a dominant peak. *J. Comput. and Graphical Stat.*, **6**, 92–111.

[34] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics* 4, (ed. J.M. Bernardo, J.O. Berger, A.P. Dawid e A.F.M. Smith). Clarendon Press, Oxford, UK.

[35] Geyer, C.J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

[36] Gillies, D. (2001). Bayesianism and the fixity of the theoretical framework. In *Foundations of Bayesianism*, J. Corfield e J. Williamson (eds.), pp. 363–379, Kluwer Academic Publishers, Dordrecht.

[37] Givens, G.H. e Hoeting, J.A. (2005). *Computational Statistics*. John Wiley & Sons.

[38] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[39] Heidelberger, P. e Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109–44.

[40] Henderson, H.V. e Velleman, P.F. (1981). Building multiple regression models interactively. *Biometrics*, **37**, 391–411.

[41] Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics*, **4**, 227–291.

[42] Jaynes, E.T. (2003). *Probability Theory*. Cambridge University Press.

[43] Kass, R.E. e Raftery, A.E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773–795.

[44] Kass, R.E. e Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.*, **91**, 1343–1370.

[45] Kempthorn, O. e Folks, L. (1971). *Probability, Statistics and Data Analysis*. The Iowa State University Press, Iowa.

[46] Korner-Nievergelt, F., von Felten, S., Roth, T., Almasi, B., Guélat, J. e Korner-Nievergelt, P. (2015). Bayesian Data Analysis in Ecology Using Linear Models with `R`, `BUGS`, and `Stan`. Academic Press.

[47] Kruschke, J. (2011). *Doing Bayesian Data Analysis: A tutorial with R and BUGS*. Academic Press/Elsevier, Amsterdam.

[48] Kruschke, J. (2014). *Doing Bayesian Data Analysis: A tutorial with R, JAGS and Stan*. Academic Press/Elsevier, Amsterdam.

[49] Lehmann, E.L. (1983). *The Theory of Point Estimation*. Wiley, New York.

[50] Lindgren, F., Rue, H. e Lindstrom, J. (2011). An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: the Stochastic Partial Differential Equation Approach. *J. Royal Statist. Soc. B*, **73(4)**, 423–498.

[51] Lindgren, F. e Rue H. (2015). Bayesian Spatial Modelling with R-INLA. To be published in *Journal of Statistical Software*.

[52] Lindley, D.V. (1990). The 1988 Wald memorial lectures: the present position in Bayesian statistics. *Statistical Science*, **5** (25), 44–89.

[53] Lunn, D., Spiegelhalter, D., Thomas, A. e Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, **28** (25), 3049–3067.

[54] MacEachern, S. e Berliner, L. (1994). Subsampling the Gibbs sampler. *The American Statistician*, **48**, 188–190.

[55] Martino, S. e Rue, H. (2009). Implementing approximate Bayesian inference using integrated nested Laplace approximation: a manual for the INLA program. *Technical Report*, Department of Mathematical Sciences, NTNU, Norway.

[56] Mayo, D. e Kruse, M. (2001). Principles of inference and their consequences. In *Foundations of Bayesianism*, J. Corfield e J. Williamson (eds.), pp. 381–403, Kluwer Academic Publishers, Dordrecht.

[57] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M. N., Teller, A.H. e Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

[58] Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Techn. Rep., Purdue University, West Lafayette, Indiana.

[59] Müller, P. (1993). Alternatives to the Gibbs sampling scheme. Techn. Rep., ISDS, Duke University.

[60] Neal, R.M. (1997). Markov Chain Monte Carlo methods based on "slicing" the density function. Techn. Rep., University of Toronto.

[61] Neal, R.M. (2003). Slice sampling (with discussion). *Ann. Statist.*, **31**, 705–767.

[62] Neal, R.M. (2011). MCMC using Hamiltonian dynamics, Chapman and Hall/CRC, ch. 5, 113–162.

[63] Newton, M. A. e Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Royal Statist. Soc. B*, **56**, 1-48.

[64] Ntzoufras, I. (2009). *Bayesian Modeling using* `WinBUGS`, Wiley: New York.

[65] O'Hagan, A. (1994). *Bayesian Inference*. Kendall's Advanced Theory of Statistics, Volume 2B, Arnold, London.

[66] O'Hagan, A. (2010). *Bayesian Inference*. 3rd ed., Kendall's Advanced Theory of Statistics, Vol. 2B, Arnold, London.

[67] Patil, V.H. (1964). The Behrens-Fisher problem and its Bayesian solution. *J. Indian Statist. Assoc.*, **2**, 21.

[68] Paulino, C.D., Soares, P. e Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics*, **59**, 670–675.

[69] Paulino, C.D., Amaral Turkman, M.A., Murteira, B. e Silva, G.L. (2018). *Estatística Bayesiana*, 2ª ed., Fundação Calouste Gulbenkian, Lisboa.

[70] Paulino, C.D. e Singer, J.M. (2006). *Análise de Dados Categorizados*. Editora Edgard Blücher, São Paulo.

[71] Plummer, M. (2003). `JAGS`: A program for analysis of Bayesian graphical models using Gibbs sampling.

[72] Plummer, M. (2012). `JAGS` version 3.3.0 user manual. http://people.math.aau.dk/~kkb/Undervisning/Bayes14/sorenh/docs/jags_user_manual.pdf

[73] Plummer, M. (2017). `JAGS: Just Another Gibbs Sampler`, version 4.3.0. Available in https://sourceforge.net/projects/mcmc-jags/files

[74] Plummer, M., Best, N.G., Cowles, M.K. e Vines, S.K. (2006). *CODA: Convergence Diagnostics and Output Analysis for MCMC. R News*, **6(1)**, 7–11. http://CRAN.R-project.org/doc/Rnews/Rnews_2006-1.pdf

[75] Raftery, A.L. e Lewis, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics* 4, (ed. J.M. Bernardo, J.O. Berger, A.P. Dawid e A.F.M. Smith), pp. 763–74. Oxford University Press.

[76] Robert, C.P. (1994). *The Bayesian Choice*. Springer-Verlag, New York.

[77] Robert, C.R. e Casella, G. (2004). *Monte Carlo Statistical Methods*. 2nd ed., Springer, New York.

[78] Robert, C.R. e Casella, G. (2010). *Introducing Monte Carlo Methods with R*. 2nd ed., Springer, New York.

[79] Ross, S.M. (2014). Introduction to Probability Models. 11th ed., Academic Press.

[80] Rue, H. e Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*, Monographs on Statistics and Applied Probability, vol. 104. Chapman & Hall, London.

[81] Rue, H., Martino, S. e Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Royal Statist. Soc. B*, **71(2)**, 319–392.

[82] Schofield, M.R., Barker, R.J., Gelman, A., Cook, E.R. e Briffa, K. (2014). Climate reconstruction using tree-ring data. *J. Amer. Statist. Assoc.* (under revision).

[83] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-466.

[84] Shaw, J.E.H. (1988). Aspects of numerical integration and summarization. In *Bayesian Statistics 3*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley e A.F.M. Smith (eds.), 625–631, University Press, Oxford.

[85] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[86] Smith, A.F.M. (1991). Bayesian computation methods. *Phil. Trans. R. Soc. Lond. A*, **337**, 369–386.

[87] Smith, A.F.M. e Gelfand, A.E. (1992). Bayesian statistics without tears. *The American Statistician*, **46**, 84–88.

[88] Smith, B. (2007). BOA: An R package for MCMC output convergence assessment and posterior inference. *J. Statist. Software* **21**(11), 1–37. http://www.jstatsoft.org/

[89] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. e van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Royal Statist. Soc. B*, **64**, 583-639.

[90] Stan Development Team (2014, a). Stan: A C++ Library for Probability and Sampling, Version 2.5.0. http://mc-stan.org

[91] Stan Development Team (2014, b). RStan: the R interface to Stan, Version 2.5. http://mc-stan.org/rstan.html

[92] Sturtz, S., Ligges, U. e Gelman, A. (2005). R2WinBUGS: A Package for Running `WinBUGS` from `R`. *Journal of Statistical Software*, **12(3)**, 1–16.

[93] Tanner, M.A. (1996). *Tools for Statistical Inference*. 3rd ed., Springer Verlag, New York.

[94] Tanner, M.A. e Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.*, **82**, 528-550.

[95] Thomas A., O' Hara, B., Ligges, U. e Sturtz, S. (2006). Making `BUGS` Open. *R News*, **6**, 12–17.

[96] Tierney, L. e Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82–86.

[97] Tierney, L., Kass, R.E. e Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. Amer. Statist. Assoc.*, **84**, 710–716.

[98] Tierney, L. (1996). Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*, 61-74, (W.R. Gilks, S. Richardson e D.J. Spiegelhalter, eds.). Chapman and Hall, London.

[99] Umlauf, N., Adler, D., Kneib, T., Lang, S. e Zeileis, A. (2015). Structured Additive Regression Models: An R Interface to BayesX. *Journal of Statistical Software*, **63**, 1–46.

[100] Vehtari, A. e Gelman, A. (2014). WAIC and cross-validation in Stan. Unpublished work.
http://www.stat.columbia.edu/~gelman/research/unpublished/waic_stan.pdf

[101] Vehtari, A. e Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, **6**, 142–228.

[102] Vehtari, A., Gelman A. e Gabry, J. (2015). Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models.
http://www.stat.columbia.edu/~gelman/research/unpublished/loo_stan.pdf

[103] Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions. *J. Royal Statist. Soc. B*, **315**, 80–88.

[104] Wasserman, L. (2004). *All of Statistics*. Springer, New York.

[105] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross-validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

Índice Remissivo

XXII CONGRESSO
SPE2015
SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

http://spe2015.mozello.com

UAlg
UNIVERSIDADE DO ALGARVE

CEAUL
Centro de Estatística e Aplicações
Universidade de Lisboa

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

ESCOLAR EDITORA

HPZ

FIDELIDADE
SEGUROS DESDE 1808

cePAC
CENTRO DE ESTUDOS EM PATRIMÓNIO
PAISAGEM E CONSTRUÇÃO

município de Olhão

MULTICOPIAS
CENTRO DE COPIAS

EDIÇÕES SÍLABO, Lda.
Publicamos conhecimento www.silabo.pt

CENTRO CONGRESSOS
CONGRESS CENTER
RIA FORMOSA

untappedevents
Organização e Gestão de Eventos
Científicos, Empresariais, Culturais e Sociais

PSE
PRODUTOS E SERVIÇOS DE ESTATÍSTICA

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

CCMAR

design: Ludovico Silva