Especificação e Estimação de Modelos Bayesianos

- Métodos de amostrar de distribuições posteriores desconhecidas
- MCMC
- Análise de saídas de MCMC

Montando o Problema em R

- Já vimos como usar métodos Monte Carlo para estimar o valor de um integral.
- Para calcular integrais com a posteriori precisamos primeiro de representar a posteriori no R
- Uma situação comum é ter uma amostra de algum modelos de amostragem f(y|\theta) e uma priori g(\theta). Nesse caso o log posteriori é dada por:

$$\log g(\theta|y) = \log g(\theta) + \sum_{i=1}^{n} \log f(y_i|\theta)$$

Amostra de uma Normal

- Para uma amostra da distribuição Normal com priori N(10,20) para a média e priori “chata” para log(\sigma) a posteriori é dada por:

$$\log g(\theta|y) = \log \phi(\mu; 10, 20) + \sum_{i=1}^{n} \log \phi(y_i; \mu, \sigma)$$

- Ver R

Transformação de Parâmetros

- É conveniente transformar os parâmetros para tomar valores na reta toda.
- Variâncias que são estritamente maiores que zero podem ser transformadas por log()
- Parâmetros que variam entre [0,1] podem ser transformadas pelo logit(p) = log(p/1-p)
- MCMC converge melhor se os parâmetros são independentes, então para regressões, pode centrar suas variáveis para diminuir correlações entre coeficientes.

Modelo Beta-Binomial

Table 5.1. Cancer mortality data. Each ordered pair represents the number of cancer deaths y_j and the number at risk n_j for an individual city in Missouri.

(0, 1083)	(0, 855)	(2, 3461)	(0, 657)	(1, 1208)	(1, 1025)
(0, 527)	(2, 1668)	(1, 583)	(3, 582)	(0, 917)	(1, 857)
(1, 680)	(1, 917)	(54, 53637)	(0, 874)	(0, 395)	(1, 581)
(3, 588)	(0, 383)				

- Se ajustar um modelo Binomial para esses dados, a variância do Binomial sub-estima a variância real dos dados (over-dispersion)
- Usa Beta-Binomial com média eta e precisão K

$$f(y_j|\eta, K) = \binom{n_j}{y_j} \frac{B(K\eta + y_j, K(1-\eta) + n_j - y_j)}{B(K\eta, K(1-\eta))}$$

Modelo Beta-Binomial

Suppose we assign the parameters the vague prior proportional to

$$g(\eta, K) \propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+K)^2}.$$

Then the posterior density of (η, K) is given, up to a proportionality constant by

$$g(\eta, K|\text{data}) \propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+K)^2} \prod_{j=1}^{20} \frac{B(K\eta + y_j, K(1-\eta) + n_j - y_j)}{B(K\eta, K(1-\eta))},$$

where $0 < \eta < 1$ and $K > 0$.

- Ver R para o cálculo do log da posteriori

Modelo Beta-Binomial

- a assimetria é típica de um parâmetro de precisão ou variância
- vamos aplicar a transformação log(K) sugerido nos slides
- Também aplicar a transformação logit(eta)
- A nova posterior é (Ver R)

$$g_1(\theta_1, \theta_2|\text{data}) = g\left(\frac{e^{\theta_1}}{1+e^{\theta_1}}, e^{\theta_2}\right)\frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1})^2}$$

Aproximações pelo Modo Posterior

- Perto do máximo a posteriori pode ser aproximada por uma normal multivariada com média igual ao modo e variância igual ao segundo derivado da posteriori no modo

$$h(\theta, y) = \log(g(\theta) f(y|\theta))$$

$$h(\theta) \approx h(\hat{\theta}) + (\theta - \hat{\theta})' h''(\hat{\theta})(\theta - \hat{\theta})/2,$$

$$V = (-h''(\hat{\theta}))^{-1}.$$

- Ver R

Sampling Importance Resampling

- O restante de capítulo 5 trata de métodos de amostragem da posteriori, que já vimos nas aulas anteriores
- Mas, tem um método novo que segue a ideia de “Rejection Sampling”, chamado “Sampling Importance Resampling”. Segue o método:
- Amostrar de uma distribuição candidata e calcula pesos iguais à razão entre a posteriori e essa distribuição
- Re-amostrar da amostra com probabilidades proporcionais aos pesos.

Markov Chain Monte Carlo (MCMC)

- Os métodos de amostragem anteriores funcionam bem se tiver uma distribuição candidata.
- Em altas dimensões pode ser difícil encontrar candidatas.
- MCMC explora a posteriori andando aleatoriamente por ela.

Cadeias de Markov Discretas

- Imagina um passeio aleatório nos números de 1 a 6.
- Se a posição for de 2 a 5 a probabilidade é 50% de mudar ou ficar, se for mudar a probabilidade é 50% de mudar uma posição à esquerda ou uma posição à direita.
- Se a posição for 1 ou 6 a probabilidade é 50% de mudar uma posição ou ficar.
- Essa cadeia tem seis estados e podemos representar as probabilidade de transição numa matriz.

Cadeias de Markov Discretas

$$P = \begin{bmatrix} .50 & .50 & 0 & 0 & 0 & 0 \\ .25 & .50 & .25 & 0 & 0 & 0 \\ 0 & .25 & .50 & .25 & 0 & 0 \\ 0 & 0 & .25 & .50 & .25 & 0 \\ 0 & 0 & 0 & .25 & .50 & .25 \\ 0 & 0 & 0 & 0 & .50 & .50 \end{bmatrix}$$

- Esta cadeia é *irreducible*, você chega a todos os estados de qualquer ponto inicial
- Esta cadeia é *aperiodic*, você pode voltar para o mesmo estado em qualquer número de passos depois

Cadeias de Markov Discretas

- Uma localização pode ser representada por um vetor de probabilidades p=(p1,p2,p3,p4,p5,p6)
- A localização no próximo passo é dado por

  p_{j+1} = p_{j} P
- Podemos falar da distribuição estacionária que satisfaz w = w P. Quando j → infinito p_{j} → w
- Se a cadeia for *irreducible* e *aperiodic* a distribuição estacionária é única.
- Ver R

MCMC Metropolis-Hastings

- Vimos que a distribuição dos valores do passeio são amostras da distribuição estacionária.
- A ideia atras de Metropolis-Hastings é simular um passeio pelo espaço paramétrico que representa uma amostra da distribuição posteriori.
- Metropolis e Hastings desenvolverem suas cadeias para ter a propriedade da distribuição estacionária ser a posteriori.

MCMC Metropolis-Hastings

- Os insumos são
  - a posteriori g(θ|y) que denotamos g(θ)
  - Um ponto inicial θ_0
  - Uma regra para gerar e aceitar ou não propostas θ*
- A regra precisa de uma densidade de proposta, p(θ*|θ_{t-1}) e uma probabilidade de aceitação

MCMC Metropolis-Hastings

- Simulate a candidate value θ^* from a proposal density $p(\theta^*|\theta^{t-1})$.
- Compute the ratio

$$R = \frac{g(\theta^*)p(\theta^{t-1}|\theta^*)}{g(\theta^{t-1})p(\theta^*|\theta^{t-1})}.$$

- Compute the acceptance probability $P = \min\{R, 1\}$.
- Sample a value θ^t such that $\theta^t = \theta^*$ with probability P; otherwise $\theta^t = \theta^{t-1}$.

- Se tiver um bom candidato, pode usar

  p(θ*|θ_{t-1})=p(θ*) (*independence chain)*
- Caso contrário pode fazer um passeio aleatória com h simétrico (*random walk chain*)

$$p(\theta^*|\theta^{t-1}) = h(\theta^* - \theta^{t-1})$$

$$R = \frac{g(\theta^*)}{g(\theta^{t-1})}$$

MCMC Metropolis-Hastings

- O pacote 'LearnBayes' oferece duas funções,
  - 'rwmetrop' com proposta $\theta^* = \theta_{t-1} + scale * Z$ com $Z \sim N(0,V)$
  - 'indepmetrop' com proposta $\theta^* \sim N(u,V)$
- A saída é uma matriz com os parâmetros simulados pela cadeia e a taxa de aceitação.
- A taxa de aceitação deve ficar em torno de 25% para mais que dois parâmetros e em torno de 45% para um e dois parâmetros. (regra de bolso)

MCMC Gibbs Sampling

- Quando o vetor de parâmetros tiver alta dimensão, podemos ter dificuldade para resumir a posteriori.
- Gibbs Sampling amostra sequencialmente das distribuições de poucos componentes, condicional nos outros componentes do vetor de parâmetros.

$$[\theta_1|\theta_2, ..., \theta_p, \text{data}],$$

$$[\theta_2|\theta_1, \theta_3, ..., \theta_p, \text{data}],$$

$$[\theta_p|\theta_1, ..., \theta_{p-1}, \text{data}],$$

MCMC Gibbs Sampling

- Se for difícil amostrar de uma das condicionais, podemos usar Metropolis-Hastings para esse.
- A função 'gibbs' faz Metropolis dentro de Gibbs.
- Depois veremos como WinBugs permite especificar as distribuições condicionais, caso essas sejam conhecidas.

Análise das saídas de MCMC

- A teoria nós diz que a cadeia gerará amostras da posteriori para comprimentos infinitos.
- Qual número de iterações é suficiente para ser “infinito”?
- Podemos analisar: a taxa de aceitação; gráficos das saídas; estatísticas resumo.
- Período de “*burn in*”. No início da cadeia os valores não saiam da posteriori. Analisa graficamente θ, componentes ou funções de θ sobre tempo (*trace plot*)

Análise de saídas de MCMC

- Autocorrelação. Se θ_{j} e θ_{j+L} tem correlação alta para L grande, estamos gerando muitos valores dependentes (pensa em amostragem de conglomerados ou clusters) assim, não estamos explorando a posteriori tanto quanto podíamos.
- Como as amostras não são independentes, não podemos calcular o erro da simulação assumindo independência. Então calcula médias de blocos e a variância dessas médias.

Uma estratégia para começar

- Precisa de uma função que calcula o log(g(θ|y))
- Precisa de um ponto inicial, e uma escala.
- Pode usar máxima verossimilhança ou o método de momentos para obter um valor inicial; ou pode usar um método de otimização para achar o modo e a variância da posteriori (função 'laplace' ou 'otim')
- Mesmo assim, eu questiono a utilidade em alta dimensão.....

Um Exemplo com Dois Parâmetros

- Temos um histograma de alturas e queremos estimar a melhor distribuição Normal que se ajuste aos dados.

Height Interval (in.)	Frequency
less than 66	14
between 66 and 68	30
between 68 and 70	49
between 70 and 72	70
between 72 and 74	33
over 74	15

- A verossimilhança é:

$$
\begin{aligned}
L(\mu,\sigma) \propto\ & \Phi(66,\mu,\sigma)^{14}(\Phi(68,\mu,\sigma)-\Phi(66,\mu,\sigma))^{30} \\
& \times (\Phi(70,\mu,\sigma)-\Phi(68,\mu,\sigma))^{49}(\Phi(72,\mu,\sigma)-\Phi(70,\mu,\sigma))^{70} \\
& \times (\Phi(74,\mu,\sigma)-\Phi(72,\mu,\sigma))^{33}(1-\Phi(74,\mu,\sigma))^{15}.
\end{aligned}
$$

Um Exemplo com Dois Parâmetros

- Para a priori proporcional a 1/σ temos:

$$g(\mu, \sigma|\text{data}) \propto \frac{1}{\sigma} L(\mu, \sigma).$$

- Transformando para λ=log(σ) temos:

$$g(\mu, \lambda|\text{data}) \propto L(\mu, \exp(\lambda))$$

- Ver R

Outros Exemplos

- O restante de capítulo 6 mostra mais dois exemplos.
- Um que compara vários métodos de estimação para uma distribuição Cauchy.
- Outro que modela a sobrevivência de pacientes que passaram por transplantes de coração.

Modelos Hierárquicos

- Podem existir modelos com muitos parâmetros.
- Se acreditar que esses parâmetros tenham uma relação entre si, podemos tentar descrever essa relação via uma única priori que descreve a distribuição de todos os parâmetros.
- Por exemplo, podemos acreditar que as taxas de mortalidade por transplantes de coração são diferentes entre hospitais, mas que todas vêm de uma mesma distribuição.

Modelos Hierárquicos

- Os dados seguem $y \sim f(y|\theta)$
- Os parâmetros, θ, seguem uma distribuição com hiper parâmetros λ que são desconhecidos $\theta \sim g_1(\theta|\lambda)$
- Os hiper parâmetros, λ, seguem uma distribuição conhecida $\lambda \sim g_2(\lambda)$
- A posteriori é $g_2 * g_1 * f$

Modelos Hierárquicos

- Estimativas individuais (λ diferente para cada hospital)
- Estimativa combinada (mesma λ para todos os hospitais; soma das mortes / soma das exposições)
- Veja no R o gráfico de taxas individuais contra log(exposição)
- As taxas individuais variam muito e podem ser estimativas pobres para exposições pequenas.

Modelos Hierárquicos

- A estimativa combinada coloca todas as taxas iguais a $\frac{\sum_{j=1}^{94} y_j}{\sum_{j=1}^{94} e_j}$
- Estimativas combinadas podem não refletir a verdadeira variação de taxas entre hospitais.
- A Shrinkage Estimate oferece uma combinação das duas estimativas:

$$(1-\lambda)\frac{y_i}{e_i} + \lambda\frac{\sum_{j=1}^{94} y_j}{\sum_{j=1}^{94} e_j}$$

Taxas Iguais

- Primeiro vamos explorar o modelo que trata todas as taxas como sendo iguias.
- Uma maneira de avaliar esse modelo é comparar as taxas observadas com a distribuição preditiva da posteriori.

$$g(\lambda) \propto \frac{1}{\lambda} \qquad g(\lambda|\text{data}) \propto \frac{1}{\lambda} \prod_{j=1}^{94} \left[\lambda^{y_j} \exp(-e_j\lambda)\right]$$

$$= \lambda^{\sum_{j=1}^{94} y_j - 1} \exp\left(-\sum_{j=1}^{94} e_j\lambda\right)$$

- Ver R.

$$f(y_i^*|e_i, y) = \int f_P(y_i^*|e_i\lambda)g(\lambda|y)d\lambda.$$

Uma priori *"Exchangable"*

- Agora assume que as λ's vem da distribuição gamma(\alpha,\mu):

$$g(\lambda|\alpha,\mu) = \frac{(\alpha/\mu)^{\alpha}\lambda^{\alpha-1}\exp(-\alpha\lambda/\mu)}{\Gamma(\alpha)}, \lambda > 0.$$

- \mu ~ gamma_inverso(a,b)
- A distribuição conjunta de λ_1 e λ_2 é

$$g(\lambda_1,\lambda_2|\alpha_0) \propto \frac{(\lambda_1\lambda_2)^{\alpha_0-1}}{(\alpha_0(\lambda_1+\lambda_2)+b)^{2\alpha_0+a}}$$

- Ver R para o gráfico

Prioris para os hiper parâmetros

- Uma priori não informativa para a média

$$g(\mu) \propto \frac{1}{\mu}, \mu > 0.$$

- Uma priori própria, mas difusa para a precisão

$$g(\alpha) = \frac{z_0}{(\alpha + z_0)^2}, \alpha > 0$$

- A posteriori é gama(y_i+alpha, e_i+alpha/mu)

$$E(\lambda_i|y, \alpha, \mu) = \frac{y_i + \alpha}{e_i + \alpha/\mu} = (1 - B_i)\frac{y_i}{e_i} + B_i\mu, \qquad B_i = \frac{\alpha}{\alpha + e_i\mu}$$

Simulando da posteriori

- Posteriori marginal de alpha e mu:

$$p(\alpha, \mu|\text{data}) = K \frac{1}{\Gamma^{94}(\alpha)} \prod_{j=1}^{94} \left[\frac{(\alpha/\mu)^{\alpha} \Gamma(\alpha + y_i)}{(\alpha/\mu + e_i)^{(\alpha+y_i)}} \right] \frac{z_0}{(\alpha + z_0)^2} \frac{1}{\mu}$$

- Posteriori conjunta =

$$g(\text{hyperparameters}|\text{data})\ g(\text{true rates}|\text{hyperparameters}, \text{data})$$

- Simula os hiper parâmetros, depois os parâmetros de cada hospital

Simulando da posteriori

- Transformam os hiper parâmetros pelo log()

$$\theta_1 = \log(\alpha), \theta_2 = \log(\mu)$$

- A nova posteriori marginal para esses parâmetros é:

$$p(\theta_1, \theta_2|\text{data}) = K \frac{1}{\Gamma^{94}(\alpha)} \prod_{j=1}^{94} \left[\frac{(\alpha/\mu)^{\alpha} \Gamma(\alpha + y_i)}{(\alpha/\mu + e_i)^{(\alpha+y_i)}} \right] \frac{z_0 \alpha}{(\alpha + z_0)^2}$$

- Ver R para simulações da posteriori

Inferência da Posteriori

- Vimos que a estimativa de uma taxa é uma média entre a taxa bruta do hospital e a média geral, isso se chama *"shrinkage".*

$$E(\lambda_i|\text{data}) \approx (1 - E(B_i|\text{data}))\frac{y_i}{e_i} + E(B_i|\text{data})\frac{\sum_{j=1}^{94} y_j}{\sum_{j=1}^{94} e_j}$$

$$B_i = \alpha/(\alpha + e_i\mu)$$

- Ver R para o gráfico de shrinkage vs exposição

Inferência da Posteriori

- Comparando hospitais
- Qual hospital tem a menor taxa média de morte?
- Qual a probabilidade da taxa de cada hospital ser menor que a taxa das outras?
- Ver R para a resposta a essas questões

Análise de Sensitividade

- Será que os resultados dependem fortemente da priori?
- Como alpha gera grandes correlações entre os lambdas, vamos mudar a priori para alpha, trocando a média de 0.53 a 5.
- $$g(\theta_1|z_0) = \frac{z_0 \exp(\theta_1)}{(z_0 + \exp(\theta_1))^2}$$
- Vamos usar SIR para reamostrar dos valores de alpha que já temos. Os pesos são:

$$w(\theta_1) = \frac{g(\theta_1|z_0 = 5)}{g(\theta_1|z_0 = 0.53)}$$

Ver R para a implementação.

Verificação da preditiva da posteriori

- No início da análise vimos que uma taxa constante não foi um bom modelo, pois tinha muitos valores com valores nas caudas da distribuição preditiva da posteriori.
- Vamos verificar para o novo modelo.
- Hospital 94 tinha 17 mortes. Isso bate com o modelo?
- Ver R