Universidade Federal de São Carlos - UFSCar
Centro de Ciências Exatas e Tecnologia
Programa de Pós-Graduação em Estatística
Departamento de Estatística

MODELO BAYESIANO DE COINCIDÊNCIAS EM PROCESSOS DE LISTAGENS

Juliana Coutinho dos Reis

Orientador: Prof. Dr. José Galvão Leite

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos
Março de 2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

R375mb

Reis, Juliana Coutinho dos.
Modelo bayesiano de coincidências em processos de listagens / Juliana Coutinho dos Reis. -- São Carlos : UFSCar, 2006.
89 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2006.

1. Cadeias de Markov. 2. Coincidência. 3. Listas. 4. Inferência bayesiana. 5. Gibbs sampling. I. Título.

CDD: 519.233 (20ª)

Agradeço,

À Deus pela força durante todos esses anos de estudo.

Aos meus pais e minha irmã pela paciência, apoio e compreensão durante todo esse processo.

Ao professor Dr. José Galvão Leite pela orientação, pelas idéias e principalmente pelo exemplo de dedicação e disciplina ao trabalho.

À CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) pela assistência financeira.

À todos os professores do Departamento de Estatística da UFSCar, em especial àqueles aos quais tive o privilégio de ser aluna.

À todos meus colegas de pós-graduação, em especial ao meu amigo Marcelo, pela convivência e amizade durante todo o período de realização do mestrado.

À minha grande amiga Lislaine e a todos que, de maneira direta ou indireta, me apoiaram em todos os momentos.

Resumo

Nesta dissertação apresentamos uma metodologia bayesiana para estimar o número de indivíduos coincidentes de duas listas, considerando a ocorrência de registros corretos e incorretos das informações cadastrais de cada indivíduo presente nas listas. Adotamos três diferentes *prioris* para o número de pares coincidentes e estudamos sua performance através de dados simulados. Devido às dificuldades encontradas na escolha dos valores dos hiperparâmetros deste modelo, apresentamos como solução a este problema um modelo bayesiano hierárquico e verificamos sua adequabilidade através das estimativas obtidas para dados simulados.

Abstract

In this work we present a bayesian methodology to estimate the number of coincident individuals of two lists, considering the occurrence of correct and incorrect registers of the informations registers of each individual present in the lists. We adopt, in this model, three different *prioris* for the number of coincident pairs and study its performance through simulated data. Due to difficulties found in the choice of the hiperparameters of this model, we present as solution to the this problem a hierarchic bayesian model and verify its adequateness through the gotten estimates for simulated data.

Sumário

Capítulo 1

Introdução

A estimação do número de indivíduos coincidentes em amostras selecionadas com reposição de uma população é um assunto de interesse crescente na Estatística. Esse é motivado, principalmente, pela necessidade em se obter mais e melhores informações sobre os indivíduos que fazem parte de um determinado estudo.

Dessa forma, o objetivo é estimar o número de indivíduos coincidentes em amostras de uma população que, nesta dissertação, serão tratadas como duas listas. A técnica adotada consiste em considerar duas listas, ou arquivos, de indivíduos de uma certa população, aos quais podemos associar um conjunto de informações como nome, sobrenome, idade, sexo, R.G., C.I.C., endereço, etc. Estas informações serão chamadas variáveis explicativas. Através da comparação dos valores destas variáveis podemos então, decidir se os indivíduos que estão sendo comparados podem ser considerados realmente o mesmo indivíduo. Esta decisão é aceita se todos os valores das variáveis observadas forem iguais e neste caso, dizemos que os indivíduos são coincidentes. Se considerarmos uma situação em que as informações sobre os indivíduos são fornecidas e registradas corretamente, então o número observado de indivíduos coincidentes é correto e não precisa ser estimado. Todavia, o problema é que nem sempre as informações obtidas estão corretas e, por esse motivo, vamos considerar a existência de erros nas informações registradas.

Vários pesquisadores têm se dedicado a este assunto, tais como Belin *et al.* (1995), Fellegi *et al.* (1974), Fienberg *et al.* (1999), Fortini *et al.* (2000) e (2001), Jaro (1989), Lee (2002), Micheletti (2003), mas, segundo nosso conhecimento, existem poucos trabalhos sob o enfoque bayesiano sobre o assunto. Em particular, um estudo nesse mesmo sentido

realizado por Micheletti (2003) apresentou estimativas para o número de coincidências sob o ponto de vista da estatística clássica. Nossa proposta nessa dissertação é então, estimar, sob um enfoque bayesiano, o número de coincidências provenientes de um processo de listagem, considerando erros nos registros das informações.

Para melhor entender esse processo, suponhamos duas listas de pacientes que fazem tratamento de uma certa doença como, por exemplo, a hipertensão nos ambulatórios de dois hospitais. Nessas listas, cada paciente é identificado através de informações como nome, sobrenome, idade, sexo, R.G., C.I.C., endereço, etc. Porém, no momento do registro dessas informações pode haver erros por parte de quem está registrando, ou por informações erradas que o paciente possa fornecer de maneira deliberada ou mesmo acidental, omitindo ou transcrevendo informações quando solicitadas. Considere por exemplo o registro do nome. Este é sujeito a erros de ortografia e de transposição de caracteres no momento da digitação, o que pode dificultar a identificação de um par verdadeiro. Por outro lado, a existência de homônimos pode levar à identificação de um par falso. Neste caso, o objetivo do modelo de coincidências em listagens é estimar o número de pacientes coincidentes em ambas as listas, através da comparação dos dados correspondentes a cada paciente. Esses dados, ou campos preenchidos, são valores de variáveis explicativas. Esse processo pode ser considerado como uma regra de decisão que, para cada par de registros, terá uma entre três possíveis ações: coincidência, possível coincidência ou não coincidência.

Como os valores associados às variáveis explicativas nem sempre informam corretamente os verdadeiros valores dos dados de cada unidade, devemos considerar os erros envolvidos.

Um exemplo típico da importância do estudo de coincidências em listagens é quando se deseja estimar o tamanho de uma população via técnica de captura-recaptura, em particular quando a população é dificilmente observável e nota-se diferenças nos valores das variáveis de identificação dos indivíduos nas várias épocas de amostragem. Outro exemplo é dado pela possibilidade de se ordenar uma base de dados administrativos ou para completar arquivos em um levantamento de dados.

Em geral, a combinação de dois (ou mais) arquivos de dados pode ser interessante quando se deseja obter uma base de dados mais ampla, em termos de registros, e mais

rica, em termos de informações que esses arquivos contém, e assim desenvolver análises estatísticas baseadas em informações que não são conhecidas originalmente.

Neste trabalho, consideramos duas listas de indivíduos de uma população, e a cada indivíduo associamos um conjunto de informações na forma de valores assumidos por variáveis explicativas. Mais especificamente, cada indivíduo será representado por dois subconjuntos de informações, e a coincidência entre dois indivíduos será aceita se um e/ou outro subconjunto coincidir em ambas as listas. Baseados no número de indivíduos cujas informações coincidiram em somente um dos subconjuntos e em ambos, determinamos estimativas de máxima verossimilhança dos parâmetros do modelo e, através de um modelo bayesiano, estimativas *a posteriori* para o parâmetro de interesse, considerando diferentes *prioris*. A título de ilustração, apresentamos exemplos com dados simulados.

No capítulo 2 desenvolvemos o modelo estatístico e a função de verossimilhança para o caso de duas listas de indivíduos. Apresentamos ainda um estudo do modelo bayesiano no qual supomos *prioris* Uniforme, Binomial e de Poisson para o número de pares coincidentes e determinamos estimativas de máxima verossimilhança para o parâmetro de interesse. Apresentamos também exemplos com dados simulados.

No capítulo 3 implementamos o modelo bayesiano, utilizando dados simulados, através de dois métodos diferentes. No primeiro, chamado método de relação recursiva, utilizamos a distribuição *a posteriori* marginal quase exata para determinar as estimativas do parâmetro de interesse. No segundo, utilizamos o método de busca em tabela estática e as distribuições condicionais dos parâmetros para a obtenção das estimativas, através do algoritmo *Gibbs sampling*.

No capítulo 4 desenvolvemos um modelo bayesiano hierárquico para a estimação do número de indivíduos coincidentes a fim de eliminar o problema encontrado no capítulo 3, sobre a escolha dos valores dos hiperparâmetros. Este modelo foi desenvolvido considerando as distribuições *a priori* Uniforme, Binomial e de Poisson para o parâmetro de interesse. Por fim, o capítulo 5 ilustra os resultados obtidos através da implementação do modelo bayesiano hierárquico via algoritmo *Gibbs sampling*.

Capítulo 2

Estimação bayesiana do número de indivíduos coincidentes: duas listas

Neste capítulo apresentamos um modelo estatístico para a estimação do número de pares de indivíduos coincidentes em duas listas, onde consideramos a possibilidade de um par ser coincidente sem que todos os dados identificadores do par sejam registrados corretamente.

Determinamos as estimativas de máxima verossimilhança e bayesianas para os parâmetros do modelo e, através de exemplos com dados simulados, analisamos a performance das estimativas de máxima verossimilhança do parâmetro de interesse.

Consideremos duas listas, A e B, de indivíduos de uma população. Denotemos por $A = \{a_{1,}a_2, ..., a_{n_A}\}$ e $B = \{b_{1,}b_2, ..., b_{n_B}\}$. Sejam

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

o conjunto dos pares de indivíduos das duas listas e

$$\mathcal{M} = \{(a, b) \in A \times B : a = b\}$$

o conjunto dos pares de indivíduos coincidentes nas duas listas. Denotemos por n_{AB} o cardinal de $\mathcal{M}$ ou o número de pares coincidentes. Evidentemente, o cardinal de A é n_A e o cardinal de B é n_B.

Identificamos cada indivíduo pertencente às listas através dos valores assumidos por um conjunto de variáveis explicativas, $\{X_1, X_2, ..., X_k\}$, cujas componentes podem ser por exemplo, X_1 o nome, X_2 o sobrenome, X_3 a data de nascimento, X_4 o sexo, X_5 o endereço, X_6 o número do C.P.F., ..., X_k o número do R.G.. Para todo $(a, b) \in A \times B$ e $r = 1, 2, ..., k$, denotemos por $\mathbf{x}_{a,r}^A$ e $\mathbf{x}_{b,r}^B$ os valores da variável X_r no indivíduo a e no indivíduo b, respectivamente. Em uma situação ideal, isto é, em uma situação em que os valores das variáveis explicativas fossem registrados sem erros e os indivíduos a e b respondessem corretamente às questões formuladas, teríamos $\mathbf{x}_{a,r}^A = \mathbf{x}_{b,r}^B$, para todo $(a, b) \in \mathcal{M}$. Contudo, como existe a possibilidade de que alguns dados sejam registrados erroneamente, bem como os indivíduos a e b podem responder deliberada ou incorretamente à algumas questões formuladas, admitimos que pode existir $r, 1 \leq r \leq k$, tal que $\mathbf{x}_{a,r}^A \neq \mathbf{x}_{b,r}^B$. Logo, n_{AB} é um parâmetro desconhecido e nosso interesse é estimá-lo. Para isto vamos construir um modelo estatístico que possibilite estimar n_{AB} levando em conta eventuais erros de registros de dados dos indivíduos.

Suponhamos então que o conjunto de variáveis explicativas seja dividido em dois subconjuntos distintos, C e D, com c e d elementos, respectivamente. Desse modo cada indivíduo de qualquer lista será identificado pelos valores assumidos por dois conjuntos de variáveis C e D que, sem perda de generalidade, denotamos por

$$C = \{X_1, X_2, ..., X_c\} \qquad \text{e} \qquad D = \{X_{c+1}, X_{c+2}, ..., X_{c+d}\},$$

onde $c + d = k$.

Admitimos que se os valores assumidos pelas variáveis do conjunto C e/ou D forem registrados corretamente para um indivíduo, então o indivíduo é identificado de modo único, e para decidir se dois indivíduos $a \in A$ e $b \in B$ são coincidentes ou não, comparamos, para cada conjunto C e D, os valores assumidos pelas respectivas variáveis explicativas nos indivíduos. Se esses valores forem iguais para pelo menos um dos conjuntos C e D, então decidimos pela coincidência dos indivíduos a e b.

Naturalmente, ao adotar tal decisão descartamos a possibilidade da ocorrência de uma coincidência casual de dois indivíduos $a \in A$ e $b \in B$, no caso em que o valor assumido por alguma variável explicativa do conjunto C ou D seja registrado incorretamente para

algum indivíduo. Isto significa que o conjunto C ou D não deve ser constituído, por exemplo, apenas pela variável explicativa número do R.G., cujo registro errado de um de seus algarismos poderia implicar em uma falsa coincidência. Vamos supor também que

(i) os registros certos e errados dos valores observados das variáveis explicativas dos conjuntos C e D são independentes entre os indivíduos e entre as listas A e B;

(ii) para cada indivíduo e para cada lista os registros certos e errados dos valores das variáveis explicativas do conjunto C ou D são independentes dos registros certos e errados dos valores observados das variáveis explicativas do outro conjunto, D ou C.

Notamos pela suposição (ii), que as constituições dos conjuntos C e D devem ser tais que as variáveis explicativas, cujos valores são mais prováveis de serem registrados erradamente porque são complicados, devem pertencer a um mesmo conjunto C ou D.

2.1 Modelo estatístico e Função de verossimilhança

Denotemos por ϕ_C (ϕ_D) a probabilidade de que os valores das variáveis explicativas do conjunto C (D), para qualquer indivíduo de ambas as listas, sejam registrados corretamente nas duas listas. Pelas hipóteses (i) e (ii) acima, segue que a probabilidade de que os valores das variáveis explicativas dos conjuntos C e D, para qualquer indivíduo de ambas as listas, sejam registrados corretamente nas duas listas é igual a $\phi_C\phi_D$.

O conjunto das trajetórias possíveis associadas a cada par de indivíduos de $\mathcal{M}$ pode ser descrito como o subconjunto

$$\{CO,CO;CO,OD;CO,CD;CO,OO;CD,CO;CD,OD;CD,CD;CD,OO;$$
$$OD,CO;OD,OD;OD,CD;OD,OO;OO,CO;OO,OD;OO,CD;OO,OO\},$$

onde

- CO,CO significa que os valores das variáveis explicativas do conjunto C foram registrados corretamente em ambas as listas e os valores das variáveis explicativas do conjunto D foram registrados incorretamente em ambas as listas;

- CO, OD significa que os valores das variáveis explicativas do conjunto C foram registrados corretamente na lista A e incorretamente na lista B e os valores das variáveis explicativas do conjunto D foram registrados incorretamente na lista A e corretamente na lista B, e assim por diante.

O problema é que o número de indivíduos que apresentam certas trajetórias não são observáveis como, por exemplo, o número de indivíduos que apresentam a trajetória CD, CO. Isto deve-se ao fato de que estes indivíduos são indistingüíveis daqueles que apresentam a trajetória CO, CO. Tudo o que sabemos é que os conjuntos D's são diferentes, mas não sabemos quais os valores das variáveis que estão incorretos. Por outro lado, o número de indivíduos que apresentam a trajetória CD, CD, por exemplo, é observável.

Denotemos por $m_{CD,CO}, m_{CD,OD}, m_{CO,CD}, m_{OD,CD}, m_{CO,CO}, m_{OD,OD}$ e m_{CD} o número de pares de $\mathcal{M}$ que apresentam as trajetórias $CD, CO; CD, OD; CO, CD; OD, CD; CO, CO$; OD, OD e CD, CD, respectivamente.

Como descartamos a possibilidade da inclusão de um indivíduo em ambas as listas no caso de os valores das variáveis dos subconjuntos C ou D assumirem valores incorretos, segue que

- $m_{CO} = m_{CD,CO} + m_{CO,CD} + m_{CO,CO}$ é o número de pares de $\mathcal{M}$ que possuem apenas os valores das variáveis do conjunto C iguais em ambas as listas;

- $m_{OD} = m_{CD,OD} + m_{OD,CD} + m_{OD,OD}$ é o número de pares de $\mathcal{M}$ que possuem apenas os valores das variáveis do conjunto D iguais em ambas as listas;

- $m_{CD} = m_{CD,CD}$ é o número de pares de $\mathcal{M}$ que possuem os valores de todas as variáveis explicativas iguais em ambas as listas;

- $m_T = m_{CO} + m_{OD} + m_{CD}$ é o número observado de pares de $\mathcal{M}$, o que implica que $n_{AB} - m_T = n_{AB} - (m_{CO} + m_{OD} + m_{CD})$ é o número (não observável) de pares de $\mathcal{M}$ para os quais nenhum dos conjuntos dos valores de variáveis explicativas é igual.

De acordo com as suposições feitas, dados n_{AB} e (ϕ_C, ϕ_D), o vetor $(m_{CO}, m_{OD}, m_{CD}, n_{AB} - m_T)$ tem distribuição Multinomial com parâmetros n_{AB} e $(\phi_C(1-\phi_D), (1-\phi_C)\phi_D,$

$\phi_C\phi_D, (1-\phi_C)(1-\phi_D))$. Logo,

$$P(m_{CO}, m_{OD}, m_{CD}, n_{AB} - m_T | n_{AB}, \phi_C, \phi_D) =$$

$$= \frac{n_{AB}!}{m_{CO}!\ m_{OD}!\ m_{CD}!\ (n_{AB}-m_T)!}\ [\phi_C(1-\phi_D)]^{m_{CO}}[(1-\phi_C)\phi_D]^{m_{OD}}[\phi_C\phi_D]^{m_{CD}} \times$$

$$\times\ [(1-\phi_C)(1-\phi_D)]^{n_{AB}-m_T} =$$

$$= \frac{n_{AB}!}{m_{CO}!\ m_{OD}!\ m_{CD}!\ (n_{AB}-m_T)!}\ \phi_C^{m_{CO}+m_{CD}}\ \ \phi_D^{m_{OD}+m_{CD}}\ (1-\phi_C)^{m_{OD}+n_{AB}-m_T} \times$$

$$\times\ (1-\phi_D)^{m_{CO}+n_{AB}-m_T} =$$

$$= \frac{n_{AB}!}{m_{CO}!\ m_{OD}!\ m_{CD}!\ (n_{AB}-m_T)!}\ \phi_C^{m_C}(1-\phi_C)^{n_{AB}-m_C}\phi_D^{m_D}(1-\phi_D)^{n_{AB}-m_D},$$

onde $m_C = m_{CO}+m_{CD}$ é o número de pares de elementos de $\mathcal{M}$ cujos valores das variáveis explicativas do conjunto C são iguais em ambas as listas e $m_D = m_{OD}+m_{CD}$ é o número de pares de elementos de $\mathcal{M}$ cujos valores das variáveis explicativas do conjunto D são iguais em ambas as listas.

Assim, a função de verossimilhança é tal que

$$L(n_{AB}, \phi_C, \phi_D | m_{CO}, m_{OD}, m_{CD}) \propto$$

$$\propto \frac{n_{AB}!}{(n_{AB}-m_T)!}\ \phi_C^{m_C}(1-\phi_C)^{n_{AB}-m_C}\phi_D^{m_D}(1-\phi_D)^{n_{AB}-m_D}, \tag{2.1}$$

$m_T \leq n_{AB} \leq M = \min\{n_A, n_B\}, 0 < \phi_C < 1, 0 < \phi_D < 1.$

2.1.1 Estimativas de máxima verossimilhança

Nesta seção determinamos os estimadores de máxima verossimilhança de ϕ_C, ϕ_D e n_{AB}, que denotamos por $\widehat{\phi}_C, \widehat{\phi}_D$ e $\widehat{n}_{AB}$, respectivamente. Para isso, consideremos o

logaritmo natural do kernel da função de verossimilhança (2.1), dado por

$$l(n_{AB},\phi_C,\phi_D) = \ln(\tfrac{n_{AB}!}{(n_{AB}-m_T)!}\phi_C^{m_C}(1-\phi_C)^{n_{AB}-m_C}\phi_D^{m_D}(1-\phi_D)^{n_{AB}-m_D}) =$$

$$= \ln\left(\tfrac{n_{AB}!}{(n_{AB}-m_T)!}\right) + m_C\ln\phi_C + (n_{AB}-m_C)\ln(1-\phi_C) + m_D\ln\phi_D + \\ + (n_{AB}-m_D)\ln(1-\phi_D).$$

Então,

$$\frac{\partial l(n_{AB},\phi_C,\phi_D)}{\partial\phi_C} = \frac{m_C}{\phi_C} - \frac{n_{AB}-m_C}{1-\phi_C} = 0 \;\Rightarrow\; \phi_C = \frac{m_C}{n_{AB}},$$

$$\frac{\partial l(n_{AB},\phi_C,\phi_D)}{\partial\phi_D} = \frac{m_D}{\phi_D} - \frac{n_{AB}-m_D}{1-\phi_D} = 0 \;\Rightarrow\; \phi_D = \frac{m_D}{n_{AB}}$$

e

$$\ln(n_{AB},\phi_C,\phi_D) = \ln(n_{AB}-1,\phi_C,\phi_D) \Rightarrow$$

$$\Rightarrow \tfrac{n_{AB}!\phi_C^{m_C}(1-\phi_C)^{n_{AB}-m_C}\phi_D^{m_D}(1-\phi_D)^{n_{AB}-m_D}}{(n_{AB}-m_T)!\phi_C^{m_C}(1-\phi_C)^{n_{AB}-C-m_C}\phi_D^{m_D}(C-\phi_D)^{n_{AB}-C-m_D}}\tfrac{(n_{AB}-1-m_T)!}{(n_{AB}-1)!} = 1$$

$$\Rightarrow \tfrac{n_{AB}}{n_{AB}-m_T}(1-\phi_C)\;(1-\phi_D) = 1$$

$$\Rightarrow \tfrac{n_{AB}-m_T}{n_{AB}} = (1-\phi_C)\;(1-\phi_D)$$

$$\Rightarrow 1-\tfrac{m_T}{n_{AB}} = (1-\phi_C)(1-\phi_D).$$

Logo, resolvendo o sistema de equações acima, temos

$$1-\frac{m_T}{n_{AB}} = \left(1-\frac{m_C}{n_{AB}}\right)\left(1-\frac{m_D}{n_{AB}}\right)$$

$$\Rightarrow (m_C+m_D-m_T)\;n_{AB} = m_C\;m_D$$

$$\Rightarrow \widehat{n}_{AB} = \frac{m_C\, m_D}{m_C + m_D - m_T} = \frac{m_C\, m_D}{m_{CD}}, \tag{2.2}$$

$$\widehat{\phi}_C = \frac{m_C}{\widehat{n}_{AB}} = \frac{m_{CD}}{m_D} \tag{2.3}$$

e

$$\widehat{\phi}_D = \frac{m_D}{\widehat{n}_{AB}} = \frac{m_{CD}}{m_C}. \tag{2.4}$$

Observamos que os estimadores de máxima verossimilhança de ϕ_C, ϕ_D e n_{AB} existem se m_C, m_D e m_{CD} forem diferentes de zero, ou seja, devemos observar ao menos um par cujos elementos do conjunto C sejam iguais em ambas as listas, um par cujos elementos do conjunto D sejam iguais em ambas as listas e um par cujos elementos dos conjuntos C e D sejam iguais em ambas as listas, respectivamente.

Para a análise do comportamento de tais estimativas utilizamos dados simulados. Em todos os exemplos apresentados abaixo os valores para n_A, n_B e n_{AB} foram obtidos da seguinte maneira. Considerando inicialmente uma população contendo 1000 indivíduos, isto é, $N = 1000$, e probabilidades θ_A e θ_B de que esses indivíduos pertençam às listas A e B, respectivamente, geramos os valores de n_A, n_B e n_{AB} através de uma distribuição multinomial com parâmetros $N, \theta_A(1-\theta_B), (1-\theta_A)\theta_B, \theta_A\theta_B, (1-\theta_A)(1-\theta_B)$. Para cada valor gerado de n_A, n_B e n_{AB}, atribuímos diferentes valores para ϕ_C e ϕ_D e geramos os valores das estatísticas m_{CO}, m_{OD}, m_{CD} e $n_{AB} - m_T$ através de uma nova distribuição multinomial com parâmetros $n_{AB}, \phi_C(1-\phi_D), (1-\phi_C)\phi_D, \phi_C\phi_D, (1-\phi_C)(1-\phi_D)$. Lembrando que $m_C = m_{CO} + m_{CD}$ e $m_D = m_{OD} + m_{CD}$, obtivemos os valores de m_C e m_D.

O programa utilizado para a geração das estatísticas m_{CO}, m_{OD}, m_{CD} e $n_{AB} - m_T$ encontra-se disponível no Apêndice A.1.

Exemplo 1. Neste exemplo apresentamos as estatísticas e as estimativas de máxima verossimilhança de n_{AB}, ϕ_C e ϕ_D, considerando $n_A = 183, n_B = 289, n_{AB} = 55$ e diversos valores para as probabilidades ϕ_C e ϕ_D. Os resultados são apresentados na tabela 1.

Tabela 1. Estatísticas e estimativas de máxima verossimilhança de ϕ_C, ϕ_D e n_{AB}.

$\boldsymbol{\phi}_C$	$\boldsymbol{\phi}_D$	$\mathbf{m}_C$	$\mathbf{m}_D$	$\mathbf{m}_{CD}$	$\mathbf{m}_T$	$\widehat{\phi}_C$	$\widehat{\phi}_D$	$\widehat{n}_{AB}$
0,10	0,20	3	11	2	12	0,1818182	0,6666667	16,5
0,15	0,30	3	21	2	22	0,0952381	0,6666667	31,5
0,40	0,75	17	46	16	47	0,3478261	0,9411765	48,875
0,50	0,60	24	39	18	45	0,4615385	0,75	52
0,65	0,80	32	52	31	53	0,5961538	0,96875	53,67742
0,80	0,55	48	32	28	52	0,875	0,5833333	54,85714
0,80	0,85	42	53	41	54	0,7735849	0,9761905	54,29268

Observando os resultados apresentados na Tabela 1, notamos que, para pequenos valores de ϕ_C e ϕ_D as estimativas de n_{AB} não são muito boas, isto é, quando as probabilidades de que os valores das variáveis explicativas dos conjuntos C e D sejam registrados corretamente em ambas listas são pequenas, não conseguimos bons resultados para $\widehat{n}_{AB}$. Conforme aumentamos os valores de ϕ_C e ϕ_D, obtemos estimativas mais próximas do verdadeiro valor do parâmetro n_{AB}.

Exemplo 2. Neste exemplo apresentamos as estatísticas e as estimativas de máxima verossimilhança para n_{AB}, ϕ_C e ϕ_D, supondo $n_A = 467, n_B = 611, n_{AB} = 297$ e variando os valores de ϕ_C e ϕ_D. A tabela 2 contém os resultados obtidos.

Tabela 2. Estatísticas e estimativas de máxima verossimilhança de ϕ_C, ϕ_D e n_{AB}.

ϕ_C	ϕ_D	m_C	m_D	m_{CD}	m_T	$\widehat{\phi}_C$	$\widehat{\phi}_D$	$\widehat{n}_{AB}$
0,10	0,20	27	54	6	75	0,1111111	0,2222222	243
0,15	0,30	42	96	15	123	0,15625	0,3571429	268,8
0,40	0,75	126	230	100	256	0,4347826	0,7936508	289,8
0,50	0,60	143	188	91	240	0,4840426	0,6363636	295,4286
0,65	0,80	190	243	155	278	0,6378601	0,8157895	297,8710
0,80	0,55	237	162	129	270	0,7962963	0,5443038	297,6279
0,80	0,85	245	262	215	292	0,8206107	0,877551	298,5581

Podemos observar pela Tabela 2 que para valores maiores de ϕ_C e ϕ_D, obtivemos estimativas mais próximas do verdadeiro valor do parâmetro n_{AB}, como no exemplo anterior.

Exemplo 3. Neste exemplo, supondo $n_A = 700, n_B = 806, n_{AB} = 572$ e variando os valores de ϕ_C e ϕ_D, obtivemos as seguintes estatísticas e estimativas de máxima verossimilhança para n_{AB}, ϕ_C e ϕ_D.

Tabela 3. Estatísticas e estimativas de máxima verossimilhança de ϕ_C, ϕ_D e n_{AB}.

ϕ_C	ϕ_D	m_C	m_D	m_{CD}	m_T	$\widehat{\phi}_C$	$\widehat{\phi}_D$	$\widehat{n}_{AB}$
0,10	0,20	50	102	15	137	0,1470588	0,3	340
0,15	0,30	85	185	40	230	0,2162162	0,4705882	393,125
0,40	0,75	228	442	185	485	0,418552	0,8114035	544,7351
0,50	0,60	267	366	178	455	0,4863388	0,6666667	549
0,65	0,80	356	471	298	529	0,6326964	0,8370787	562,6711
0,80	0,55	272	339	167	444	0,4926254	0,6139706	552,1437
0,80	0,85	465	500	412	553	0,824	0,8860215	564,3204

Como nos outros exemplos, observamos que quanto maiores as probabilidades de que os valores das variáveis explicativas dos conjuntos C e D sejam registrados corretamente, melhores as estimativas obtidas para o parâmetro n_{AB}.

2.2 Modelo bayesiano

Nesta seção apresentamos o modelo bayesiano para estimar n_{AB}.

Suponhamos uma distribuição *a priori* para n_{AB},

$$\pi(n_{AB}), \qquad n_{AB} = 0, 1, \ldots, M = \min\{n_A, n_B\}.$$

Suponhamos ainda que n_{AB}, ϕ_C e ϕ_D sejam independentes *a priori*, onde ϕ_C e ϕ_D têm distribuição Beta com parâmetros α e β, α' e β', respectivamente, com α, β, α' e β' conhecidos. Então, a distribuição *a priori* conjunta de n_{AB}, ϕ_C e ϕ_D é tal que

$$\pi(n_{AB}, \phi_C, \phi_D) = \pi(n_{AB})\pi(\phi_C)\pi(\phi_D) \propto$$

$$\propto \pi(n_{AB})\ \phi_C^{\alpha-1}\ (1-\phi_C)^{\beta-1}\ \phi_D^{\alpha'-1}\ (1-\phi_D)^{\beta'-1}, \tag{2.5}$$

$0 \leq n_{AB} \leq M, 0 < \phi_C < 1, 0 < \phi_D < 1$.

Denotemos por D= (m_C, m_D, m_{CD}) o vetor de dados amostrais. Então, de (2.1) e (2.5) segue que a distribuição *a posteriori* conjunta de n_{AB}, ϕ_C e ϕ_D é tal que

$$\pi(n_{AB}, \phi_C, \phi_D|D) \propto$$

$$\propto L(n_{AB}, \phi_C, \phi_D|D)\pi(n_{AB}, \phi_C, \phi_D) \propto$$

$$\propto \frac{n_{AB}!}{(n_{AB}-m_T)!}\ \phi_C^{m_C}(1-\phi_C)^{n_{AB}-m_C}\phi_D^{m_D}(1-\phi_D)^{n_{AB}-m_D}\pi(n_{AB})\ \phi_C^{\alpha-1}\times$$

$$\times\ (1-\phi_C)^{\beta-1}\phi_D^{\alpha'-1}\ (1-\phi_D)^{\beta'-1} =$$

$$= \frac{n_{AB}!}{(n_{AB}-m_T)!}\ \pi(n_{AB})\ \ \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}\phi_D^{m_D+\alpha'-1}\times \tag{2.6}$$

$$\times\ (1-\phi_D)^{n_{AB}-m_D+\beta'-1},$$

$m_T \leq n_{AB} \leq M,\ \ 0<\phi_C<1,\ \ 0<\phi_D<1$, o que implica que as respectivas distribuições condicionais de n_{AB}, ϕ_C e ϕ_D, necessárias para a obtenção das estimativas *a posteriori* dos parâmetros através da implementação do algoritmo *Gibbs sampling,* cujos resultados serão apresentados mais adiante, e obtidas através da distribuição (2.6) são, respectivamente

- $\pi(n_{AB}|\phi_C,\phi_D,D) \propto \frac{n_{AB}!}{(n_{AB}-m_T)!}\ \pi(n_{AB})\ \ (1-\phi_C)^{n_{AB}}\ \ (1-\phi_D)^{n_{AB}} \propto$

  $$\propto \binom{n_{AB}}{m_T}\pi(n_{AB})((1-\phi_C)(1-\phi_D))^{n_{AB}},\quad m_T \leq n_{AB} \leq M; \tag{2.7}$$

- $$\pi(\phi_C|n_{AB},\phi_D,D) \propto \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1},\quad 0<\phi_C<1; \tag{2.8}$$

- $$\pi(\phi_D|n_{AB},\phi_C,D) \propto \phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1},\quad 0<\phi_D<1. \tag{2.9}$$

Notamos através das distribuições acima que ϕ_C e ϕ_D apresentam distribuições Beta com parâmetros $(m_C+\alpha, n_{AB}-m_C+\beta)$ e $(m_D+\alpha', n_{AB}-m_D+\beta')$, respectivamente.

A seguir, determinamos a distribuição *a posteriori* marginal de n_{AB} através de uma técnica de relação recursiva, que permite obter estimativas quase exatas para n_{AB}.

De (2.6) segue que

$$\pi(n_{AB}|D) \propto \binom{n_{AB}}{m_T} \pi(n_{AB}) \int \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1} \; d\phi_C \times$$

$$\times \int \phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1} \; d\phi_D =$$

$$= \binom{n_{AB}}{m_T} \pi(n_{AB}) B(m_C+\alpha, n_{AB}-m_C+\beta) B(m_D+\alpha', n_{AB}-m_D+\beta') =$$

$$= \binom{n_{AB}}{m_T} \pi(n_{AB}) \frac{\Gamma(m_C+\alpha)\;\Gamma(n_{AB}-m_C+\beta)}{\Gamma(n_{AB}+\alpha+\beta)} \;\; \frac{\Gamma(m_D+\alpha')\;\Gamma(n_{AB}-m_D+\beta')}{\Gamma(n_{AB}+\alpha'+\beta')} \propto$$

$$\propto \binom{n_{AB}}{m_T} \pi(n_{AB}) \frac{\Gamma(n_{AB}-m_C+\beta)\Gamma(n_{AB}-m_D+\beta')}{\Gamma(n_{AB}+\alpha+\beta)\Gamma(n_{AB}+\alpha'+\beta')}, \quad m_T \leq n_{AB} \leq M, \qquad (2.10)$$

onde $B(\cdot,\cdot)$ denota a função Beta e $\Gamma(\cdot)$ a função Gamma.

As distribuições descritas acima no desenvolvimento do modelo bayesiano foram apresentadas supondo uma distribuição *a priori* genérica $\pi(n_{AB})$ para o parâmetro de interesse. A seguir, vamos supor que n_{AB} tenha distribuições Uniforme, Binomial e Poisson truncada, respectivamente, e em cada caso vamos obter as distribuições de interesse para a implementação do modelo. Posteriormente faremos uma análise de como se comporta o modelo para cada *priori* através de dados simulados e reais.

2.2.1 Priori Uniforme para o número de pares coincidentes

Suponhamos que n_{AB} tenha distribuição uniforme no conjunto $\{0, 1, \ldots, M\}$, isto é,

$$\pi(n_{AB}) = \frac{1}{M+1}, \qquad 0 \leq n_{AB} \leq M.$$

Então, de acordo com (2.6), a distribuição *a posteriori* conjunta de n_{AB}, ϕ_C e ϕ_D é tal

que

$$\pi(n_{AB}, \phi_C, \phi_D|D) \propto$$

$$\propto \frac{n_{AB}!}{(n_{AB}-m_T)!}\left(\frac{1}{M+1}\right)\ \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}\phi_D^{m_D+\alpha'-1}\times$$

$$\times\,(1-\phi_D)^{n_{AB}-m_D+\beta'-1}, \qquad (2.11)$$

$m_T \leq n_{AB} \leq M,\ 0 < \phi_C < 1,\ 0 < \phi_D < 1$, e de (2.7), (2.8) e (2.9) segue que as distribuições condicionais para n_{AB}, ϕ_C e ϕ_D são tais que

- $\pi(n_{AB}|\phi_C, \phi_D, D) \propto \binom{n_{AB}}{m_T}[(1-\phi_C)(1-\phi_D)]^{n_{AB}}; m_T \leq n_{AB} \leq M;$ (2.12)
- $\pi(\phi_C|n_{AB}, \phi_D, D) \propto \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1},\ 0 < \phi_C < 1;$ (2.13)
- $\pi(\phi_D|n_{AB}, \phi_C, D) \propto \phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1},\ 0 < \phi_D < 1.$ (2.14)

Observamos das distribuições (2.13) e (2.14) que ϕ_C e ϕ_D apresentam distribuições Beta com parâmetros $(m_C+\alpha, n_{AB}-m_C+\beta)$ e $(m_D+\alpha', n_{AB}-m_D+\beta')$, respectivamente. As distribuições obtidas acima serão utilizadas na implementação do modelo bayesiano através do algoritmo *Gibbs sampling,* como descrito no capítulo 3.

Podemos obter ainda, de acordo com (2.10), a distribuição *a posteriori* marginal de n_{AB}, que será utilizada na obtenção das estimativas *a posteriori* de n_{AB} de acordo com a técnica detalhada na seção 3.1, ou seja,

$$\pi(n_{AB}|D) \propto \binom{n_{AB}}{m_T}\frac{\Gamma(n_{AB}-m_C+\beta)\Gamma(n_{AB}-m_D+\beta')}{\Gamma(n_{AB}+\alpha+\beta)\Gamma(n_{AB}+\alpha'+\beta')}, \qquad (2.15)$$

$m_T \leq n_{AB} \leq M$.

2.2.2 Priori Binomial para o número de pares coincidentes

Suponhamos agora, que n_{AB} tenha distribuição Binomial com parâmetros M e p conhecido. Então,

$$\pi(n_{AB}) \propto \binom{M}{n_{AB}} p^{n_{AB}}(1-p)^{M-n_{AB}}, \qquad n_{AB} = 0, 1, ..., M.$$

De (2.6) segue que a distribuição *a posteriori* conjunta de n_{AB}, ϕ_C e ϕ_D é tal que

$$\pi(n_{AB}, \phi_C, \phi_D|D) \propto \frac{n_{AB}!}{(n_{AB}-m_T)!}\binom{M}{n_{AB}}\left(\frac{p}{1-p}\right)^{n_{AB}} \times$$

$$\times \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}\phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1} \propto$$

$$\propto \binom{n_{AB}}{m_T}\binom{M}{n_{AB}}\left(\frac{p}{1-p}\right)^{n_{AB}} \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}\phi_D^{m_D+\alpha'-1} \times$$

$$\times (1-\phi_D)^{n_{AB}-m_D+\beta'-1}, \tag{2.16}$$

$m_T \leq n_{AB} \leq M, 0 < \phi_C < 1, 0 < \phi_D < 1$, e de (2.7) temos

$$\pi(n_{AB}|\phi_C, \phi_D, D) \propto \binom{n_{AB}}{m_T}\binom{M}{n_{AB}}\left(\frac{p}{1-p}\right)^{n_{AB}} (1-\phi_C)^{n_{AB}}(1-\phi_D)^{n_{AB}} =$$

$$= \binom{n_{AB}}{m_T}\binom{M}{n_{AB}}\left[\frac{p(1-\phi_C)(1-\phi_D)}{1-p}\right]^{n_{AB}}, \quad m_T \leq n_{AB} \leq M. \tag{2.17}$$

Observamos que as distribuições condicionais de ϕ_C e ϕ_D são as mesmas obtidas em (2.8) e (2.9) e, juntamente com a condicional de n_{AB}, acima, serão utilizadas na imple-

mentação do modelo bayesiano via algoritmo de *Gibbs sampling*.

Segue ainda, de (2.10), que a distribuição *a posteriori* marginal de n_{AB}, importante na obtenção das estimativas *a posteriori* para n_{AB}, é tal que

$$\pi(n_{AB}|D) \propto \binom{n_{AB}}{m_T} \binom{M}{n_{AB}} \left(\tfrac{p}{1-p}\right)^{n_{AB}} \tfrac{\Gamma(n_{AB}-m_C+\beta)\Gamma(n_{AB}-m_D+\beta')}{\Gamma(n_{AB}+\alpha+\beta)\Gamma(n_{AB}+\alpha'+\beta')}, \tag{2.18}$$

$m_T \leq n_{AB} \leq M$.

2.2.3 Priori de Poisson para o número de pares coincidentes

Suponhamos que n_{AB} tenha distribuição *a priori* de Poisson truncada em $\{M+1, M+2, ...\}$ com parâmetro λ, λ conhecido. Então,

$$\pi(n_{AB}) = \frac{e^{-\lambda}\lambda^{n_{AB}}}{n_{AB}! \sum_{j=0}^{M} \frac{e^{-\lambda}\lambda^j}{j!}} = \frac{\lambda^{n_{AB}}}{n_{AB}! \sum_{j=0}^{M} \frac{\lambda^j}{j!}}, \qquad n_{AB} = 0, 1, \ldots, M.$$

Neste caso, segue de (2.6) que a distribuição *a posteriori* conjunta de n_{AB}, ϕ_C e ϕ_D é tal que

$$\begin{aligned} &\pi(n_{AB}, \phi_C, \phi_D|D) \propto \\ &\propto \binom{n_{AB}}{m_T} \tfrac{\lambda^{n_{AB}}}{n_{AB}!}\ \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}\ \phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}, \end{aligned} \tag{2.19}$$

$m_T \leq n_{AB} \leq M, 0 < \phi_C < 1, 0 < \phi_D < 1$, e de (2.7) segue que

$$\pi(n_{AB}|\phi_C, \phi_D, D) \propto \binom{n_{AB}}{m_T} \frac{\lambda^{n_{AB}}}{n_{AB}!}[(1-\phi_C)(1-\phi_D)]^{n_{AB}}, \tag{2.20}$$

$m_T \leq n_{AB} \leq M$, e será usada juntamente com as distribuições condicionais de ϕ_C e ϕ_D, dadas em (2.8) e (2.9), na implementação do modelo bayesiano através do algoritmo de *Gibbs sampling*.

De (2.10) temos que a distribuição *a posteriori* marginal de n_{AB} é tal que

$$\pi(n_{AB}|D) \propto \binom{n_{AB}}{m_T} \frac{\lambda^{n_{AB}}}{n_{AB}!} \frac{\Gamma(n_{AB} - m_C + \beta)\Gamma(n_{AB} - m_D + \beta')}{\Gamma(n_{AB} + \alpha + \beta)\Gamma(n_{AB} + \alpha' + \beta')}, \tag{2.21}$$

$m_T \leq n_{AB} \leq M$. Esta distribuição é necessária para a implementação do modelo bayesiano via relação recursiva, descrita na seção 3.1.

Capítulo 3

Implementação do modelo bayesiano para duas listas

Neste capítulo apresentamos as estimativas *a posteriori* para os parâmetros do modelo descrito no capítulo anterior, obtidas através de dois métodos diferentes. No primeiro método, utilizamos a distribuição *a posteriori* marginal exata de n_{AB}, o que nos proporciona resultados mais precisos. Como uma alternativa, e até mesmo a critério de comparação, apresentamos as estimativas obtidas através do uso do algoritmo *Gibbs sampling.* Nos dois casos utilizamos dados simulados e a implementação foi feita considerando os três tipos de *prioris* para n_{AB} estudadas anteriormente.

3.1 Implementação via relação recursiva

A seguir, descrevemos uma técnica para a implementação do modelo através da qual conseguimos obter a distribuição *a posteriori* marginal exata para n_{AB} e dela, os resumos *a posteriori* dos parâmetros. Este procedimento tende a ser mais preciso do que o algoritmo de *Gibbs sampling*, que normalmente é utilizado e será descrito mais adiante, pois deixamos de trabalhar com proporcionalidade e passamos a usar a distribuição exata, uma vez que esse método permite o cálculo da constante normalizadora da distribuição *a posteriori.*

Esse método foi implementado no software *Maple 7.0* e os programas utilizados seguem nos Apêndices B.1, B.2 e B.3.

A descrição do procedimento utilizado neste método segue abaixo.

Suponhamos que a distribuição *a posteriori* marginal para n_{AB} seja

$$\pi(n_{AB}|D) \propto f(n_{AB}),$$

ou seja,

$$\pi(n_{AB}|D) = kf(n_{AB}),$$

$k \in \mathbb{R}_+$, onde $f(\cdot)$ é uma função definida nos inteiros não negativos a valores em $\mathbb{R}_+$: conjunto dos números reais estritamente positivos.

Para todos os valores de n_{AB} tais que $m_T \leq n_{AB} \leq M-1$, seja

$$Q(n_{AB}) = \frac{f(n_{AB}+1)}{f(n_{AB})}.$$

Dessa forma, obtemos a relação recursiva

$$f(n_{AB}+1) = f(n_{AB})Q(n_{AB}), \tag{3.1}$$

$m_T \leq n_{AB} \leq M-1$.

Considerando m_T o valor inicial de n_{AB} e substituindo em $f(n_{AB})$, obtemos $f(m_T)$. Conseqüentemente,

$$\pi(m_T|D) = kf(m_T) \tag{3.2}$$

e utilizando $f(m_T)$ e a relação (3.1) , temos

$$\begin{aligned}
\pi(m_T+1|D) &= kf(m_T+1) = kf(m_T)Q(m_T), \\
\pi(m_T+2|D) &= kf(m_T+2) = kf(m_T+1)Q(m_T+1), \\
&\ldots \\
\pi(M|D) &= kf(M-1)Q(M-1).
\end{aligned} \tag{3.3}$$

Somando (3.2) e as relações (3.3) membro a membro, podemos então obter o valor

exato da constante k de tal forma que

$$1 = \sum_{j=m_T}^{M} kf(j) \Longrightarrow k = \frac{1}{\sum\limits_{j=m_T}^{M} f(j)}.$$

Uma vez conhecido o valor de k, substituimos seu valor em (3.2) e nas relações (3.3) e obtemos o valor exato de $\pi(n_{AB}|D)$, $m_T \leq n_{AB} \leq M$.

Além disso, a média e a variância *a posteriori* de n_{AB} são dadas por

$$E(n_{AB}|D) = \sum_j j\pi(j|D)$$

e

$$Var(n_{AB}|D) = \sum_j [j - E(n_{AB}|D)]^2 \pi(j|D).$$

A seguir, apresentamos exemplos com dados simulados da implementação do modelo através do método de relação recursiva, considerando as *prioris* para n_{AB} estudadas no capítulo anterior. Nos três casos de *priori* para n_{AB}, assumimos $n_A = 467, n_B = 611$ e $n_{AB} = 297$, atribuímos a (ϕ_C, ϕ_D) os valores $(0,10; 0,20)$, $(0,50; 0,60)$ e $(0,80; 0,85)$, respectivamente, e em cada caso consideramos diferentes valores para os hiperparâmetros α, β, α' e β'.

3.1.1 Priori Uniforme

No caso da distribuição *a priori* Uniforme para n_{AB}, segue de (2.15) e (3.1) que as funções $f(n_{AB})$, $f(n_{AB}+1)$ e $Q(n_{AB})$, considerando as devidas simplificações, são dadas por

$$f(n_{AB}) = \frac{n_{AB}!}{(n_{AB} - m_T)!} \frac{\Gamma(n_{AB} - m_C + \beta)\Gamma(n_{AB} - m_D + \beta')}{\Gamma(n_{AB} + \alpha + \beta)\Gamma(n_{AB} + \alpha' + \beta')}, \tag{3.4}$$

$$f(n_{AB}+1) = \frac{(n_{AB}+1)!}{(n_{AB}-m_T+1)!}\frac{\Gamma(n_{AB}-m_C+\beta+1)\Gamma(n_{AB}-m_D+\beta'+1)}{\Gamma(n_{AB}+\alpha+\beta+1)\Gamma(n_{AB}+\alpha'+\beta'+1)} \tag{3.5}$$

e

$$Q(n_{AB}) = \frac{(n_{AB}+1)(n_{AB}-m_C+\beta)(n_{AB}-m_D+\beta')}{(n_{AB}-m_T+1)(n_{AB}+\alpha+\beta)(n_{AB}+\alpha'+\beta')}. \tag{3.6}$$

Exemplo 4. Neste exemplo consideramos $\phi_C = 0,10$, $\phi_D = 0,20$ e obtivemos, por simulação, as estatísticas $m_C = 27, m_D = 54, m_{CD} = 6$ e $m_T = 75$. Os resumos da distribuição de probabilidades *a posteriori* de n_{AB} são dados na tabela 4.

Tabela 4. Resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,10$ e $\phi_D = 0,20$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	259,9088548	200	246	308	214	76,77904820	(143, 434)
1	1	1	1	243,1836647	188	229	284	200	72,05342301	(138, 417)
0	1	0	1	286,8586472	223	277	344	242	80,08145785	(156, 448)
1	0	1	0	234,1044614	181	219	272	192	70,45094316	(133, 408)
0,3	0,5	0,8	0,6	258,3466120	199	244	305	212	76,36955490	(143, 432)
0,8	3	2	0,4	229,1904547	178	215	265	190	67,42048786	(133, 398)
5	7	10	8	154,0429381	133	149	168	142	27,94168610	(111, 219)
15	18	20	31	143,7509860	129	141	154	138	19,32087836	(111, 187)
10	50	10	50	262,2053510	224	255	291	243	51,28107141	(179, 381)
50	10	50	10	87,56725957	83	86	89	86	4,749813610	(79, 97)
40	15	30	9	94,84717135	89	93	98	93	6,787679285	(83, 109)

M: Média; Med: Mediana; Mod: Moda; DP: Desvio padrão; IC: Intervalo de credibilidade de 95%.

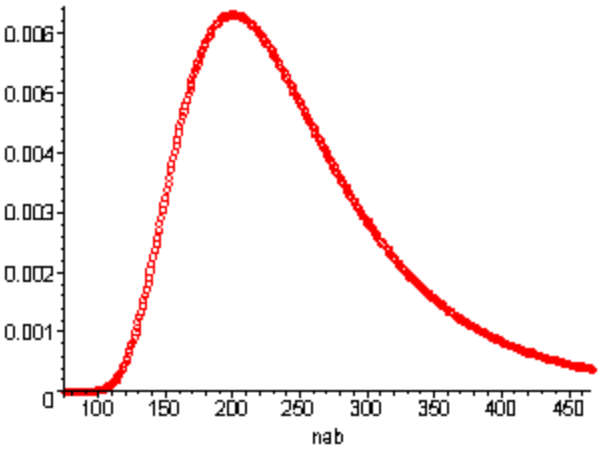

Figura 1. Gráfico da função de probabilidades *a posteriori* de n_{AB} para $\alpha = \beta = \alpha' = \beta' = 1$.

Pela tabela acima notamos que, para valores pequenos de ϕ_C e ϕ_D apenas algumas das estimativas bayesianas de n_{AB} obtidas estão próximas do verdadeiro valor do parâmetro. Em outros casos porém, não conseguimos determinar intervalos de credibilidade contendo o verdadeiro valor do parâmetro. Este fato foi observado principalmente quando os hiperparâmetros do modelo assumem valores grandes. Em particular, valores de α e α' maiores do que os valores de β e β', respectivamente, produziram estimativas inferiores em relação aos outros casos. Sendo assim, notamos a influência dos valores dos hiperparâmetros do modelo nas estimativas de n_{AB}.

A figura 1 apresenta o comportamento da função de probabilidades *a posteriori* de n_{AB} para $\alpha = \beta = \alpha' = \beta' = 1$. Para os outros casos o comportamento foi semelhante, inclusive considerando as outras *prioris* discutidas.

Exemplo 5. Consideramos $\phi_C = 0,50$, $\phi_D = 0,60$ e obtivemos, por simulação, as estatísticas $m_C = 143, m_D = 188, m_{CD} = 91$ e $m_T = 240$. Os resumos da função de probabilidades *a posteriori* de n_{AB} são dados na tabela 5.

Tabela 5. Resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,50$ e $\phi_D = 0,60$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	296,8750005	286	295	304	294	13,77376013	(272, 326)
1	1	1	1	297,0748296	286	295	305	294	13,77168126	(272, 326)
0	1	0	1	298,3595508	287	296	306	295	14,06645319	(273, 328)
1	0	1	0	295,4285713	285	293	303	292	13,48875403	(271, 324)
0,3	0,5	0,8	0,6	296,9202580	286	295	304	294	13,77720227	(272, 326)
0,8	3	2	0,4	297,0724801	286	295	305	294	13,76050999	(272, 326)
5	7	10	8	299,6450771	289	298	307	296	13,75450260	(275, 329)
15	18	20	31	313,9592812	302	312	323	312	15,29127672	(286, 346)
10	50	10	50	366,5824416	349	364	380	362	23,33895326	(324, 416)
50	10	50	10	272,8947829	266	271	277	271	8,165361754	(258, 289)
40	15	30	9	281,0949299	273	279	286	279	9,757604883	(263, 301)

M: Média; Med: Mediana; Mod: Moda; DP: Desvio padrão; IC: Intervalo de credibilidade de 95%.

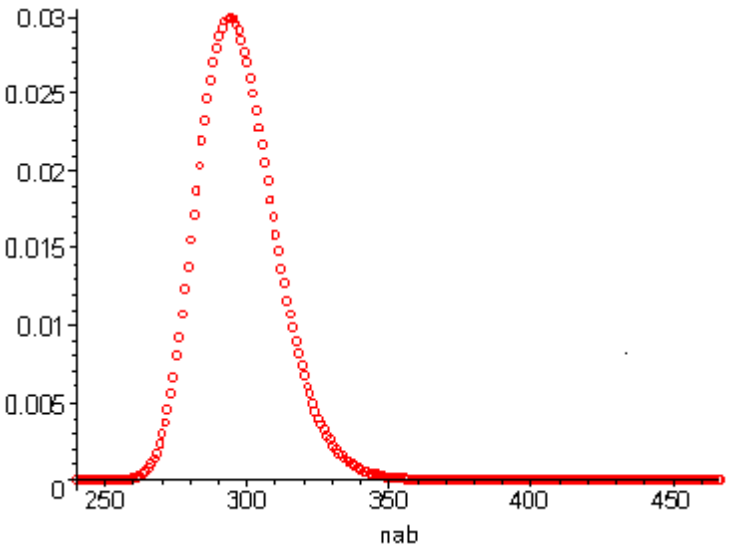

Figura 2. Gráfico da função de probabilidades *a posteriori* de n_{AB} para $\alpha = \beta = \alpha' = \beta' = 1$.

Analisando a tabela 5 observamos que para $\phi_C = 0,50$ e $\phi_D = 0,60$ conseguimos obter

boas estimativas para o parâmetro de interesse n_{AB}. Novamente observamos a influência dos valores dos hiperparâmetros nas estimativas bayesianas de n_{AB}. Notamos ainda que no caso em que $\alpha = \alpha' = 50$ e $\beta = \beta' = 10$, o intervalo de credibilidade não contém o verdadeiro valor do parâmetro.

A figura 2 apresenta o comportamento da função de probabilidades *a posteriori* de n_{AB} para $\alpha = \alpha' = \beta = \beta' = 1$. Para os outros valores o comportamento foi semelhante.

Exemplo 6. Neste exemplo atribuímos a ϕ_C o valor $0,80$, a ϕ_D o valor $0,85$ e obtivemos, por simulação, as estatísticas $m_C = 245, m_D = 262, m_{CD} = 215$ e $m_T = 292$. Os resumos da distribuição de probabilidades *a posteriori* de n_{AB} são apresentados na tabela 6.

Tabela 6. Resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,80$ e $\phi_D = 0,85$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	298,7956222	296	298	300	298	3,038149953	(293, 305)
1	1	1	1	298,9208229	296	298	300	298	3,119587796	(293, 305)
0	1	0	1	298,9859154	296	298	300	298	3,138839424	(293, 305)
1	0	1	0	298,5581398	295	297	299	298	3,024808590	(293, 304)
0,3	0,5	0,8	0,6	298,7888920	296	297	300	298	3,086280029	(293, 305)
0,8	3	2	0,4	299,0431334	296	298	300	298	3,153245428	(293, 305)
5	7	10	8	300,9756541	297	300	302	300	3,605267406	(294, 308)
15	18	20	31	307,6610693	303	306	310	307	4,954527332	(298, 317)
10	50	10	50	323,8432920	317	322	328	322	7,722720910	(309,339)
50	10	50	10	299,0242227	296	298	300	298	3,051583697	(293,305)
40	15	30	9	300,3436347	297	299	301	299	3,396967696	(294, 307)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio padrão; IC: Intervalo de credibilidade de 95%.

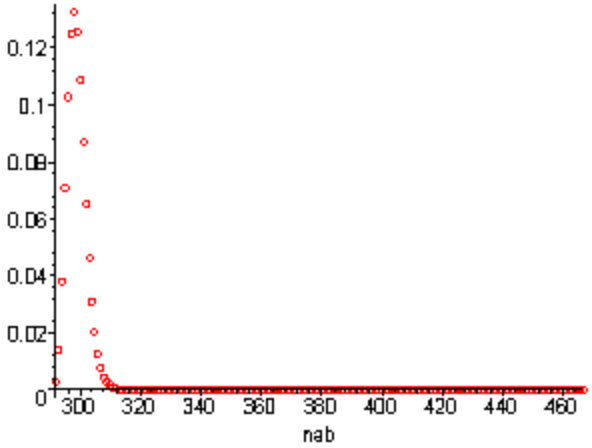

Figura 3. Gráfico da função de probabilidades *a posteriori* de n_{AB} para $\alpha = \beta = \alpha' = \beta' = 1$.

Como nos outros casos, a tabela 6 mostra uma certa influência dos valores dos hiperparâmetros do modelo nas estimativas de n_{AB}, ainda que em menor escala. Neste caso, nosso estudo mostrou que considerando grandes valores para as probabilidades ϕ_C e ϕ_D obtemos estimativas melhores em relação aos outros valores considerados.

A figura 3 descreve o comportamento da função de probabilidades *a posteriori* de n_{AB} para $\alpha = \beta = \alpha' = \beta' = 1$.

3.1.2 Priori Binomial

No caso da distribuição *a priori* Binomial para n_{AB}, as relações (2.18) e (3.1) implicam que as funções $f(n_{AB})$, $f(n_{AB}+1)$ e $Q(n_{AB})$, considerando as devidas simplificações, são dadas por

$$f(n_{AB}) = \frac{1}{(n_{AB}-m_T)!(M-n_{AB})!}\left(\frac{p}{1-p}\right)^{n_{AB}} \frac{\Gamma(n_{AB}-m_C+\beta)\Gamma(n_{AB}-m_D+\beta')}{\Gamma(n_{AB}+\alpha+\beta)\Gamma(n_{AB}+\alpha'+\beta')}; \tag{3.7}$$

$$\begin{aligned} f(n_{AB}+1) &= \frac{1}{(n_{AB}-m_T+1)!(M-n_{AB}-1)!}\left(\frac{p}{1-p}\right)^{n_{AB}+1} \times \\ &\times \frac{\Gamma(n_{AB}-m_C+\beta+1)\Gamma(n_{AB}-m_D+\beta'+1)}{\Gamma(n_{AB}+\alpha+\beta+1)\Gamma(n_{AB}+\alpha'+\beta'+1)} \end{aligned} \tag{3.8}$$

e

$$Q(n_{AB}) = \frac{(M - n_{AB})p(n_{AB} - m_C + \beta)(n_{AB} - m_D + \beta')}{(n_{AB} - m_T + 1)(1 - p)(n_{AB} + \alpha + \beta)(n_{AB} + \alpha' + \beta')}. \tag{3.9}$$

Neste caso precisamos verificar a influência do valor de p nas estimativas bayesianas de n_{AB} e, caso tal influência seja observada devemos determinar o valor de p que nos fornece as melhores estimativas. Sendo assim, para cada caso apresentado o estudo foi feito considerando valores de p variando entre $0,1$ e $0,9$. Pudemos observar com este estudo que para valores pequenos de p como $0,1$ e $0,2$, em alguns casos não obtivemos limite inferior para o intervalo de credibilidade e para valores muito altos as estimativas ultrapassaram consideravelmente o valor verdadeiro do parâmetro. Em todos os casos os valores de p que melhor estimaram o parâmetro de interesse estão entre $0,6$ e $0,7$, como mostram as tabelas abaixo. A escolha do valor ideal de p como foi feita, é trabalhosa e não muito prática. Para amenizar esse problema desenvolvemos no capítulo 4 um modelo hierárquico, atribuindo a p uma distribuição de probabilidades.

Exemplo 7. Como no exemplo 4, assumimos que $\phi_C = 0,10$, $\phi_D = 0,20$ e consideramos as estatísticas $m_C = 27, m_D = 54, m_{CD} = 6$ e $m_T = 75$. Em particular, o valor ideal para p neste caso é $0,64$ como mostra a tabela 7, que contém os resumos da função de probabilidades *a posteriori* de n_{AB}.

Tabela 7. Resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,10$, $\phi_D = 0,20$ e $p = 0,64$.

α	α'	β	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	297,9275091	290	297	304	298	10,36495131	(276, 317)
1	1	1	1	297,6261726	290	297	304	298	10,37279409	(276,317)
0	0	1	1	298,3455860	290	297	304	298	10,35116467	(277,317)
1	1	0	0	297,5067979	290	297	304	298	10,37279409	(276, 317)
0,3	0,8	0,5	0,6	297,8996477	290	297	304	298	10,36561566	(276, 317)
0,8	2	3	0,4	297,3622020	289	296	303	297	10,38037178	(276, 317)
5	10	7	8	293,8236244	286	293	300	294	10,46772983	(272, 313)
15	20	18	31	289,5295864	281	289	296	290	10,53524567	(268, 309)
10	10	50	50	297,0704402	289	296	303	297	10,28776640	(276, 316)
50	50	10	10	262,2014505	254	261	269	261	11,35566890	(239, 283)
40	30	15	9	274,3264623	266	273	281	274	11,00954822	(252, 295)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio padrão; IC : Intervalo de credibilidade de 95%.

Observamos pela tabela 7, que as estimativas bayesianas de n_{AB} são boas na maioria dos casos. Porém, em alguns casos ou para determinados valores de α, β, α' e β', como observado nas duas últimas linhas da tabela, os resultados são mais distantes do verdadeiro valor do parâmetro e respectivos intervalos de credibilidade não contém o verdadeiro valor do parâmetro.

Exemplo 8. Neste exemplo assumimos $\phi_C = 0,50$, $\phi_D = 0,60$ e consideramos as mesmas estatísticas obtidas no exemplo 5, isto é, $m_C = 143, m_D = 188, m_{CD} = 91$ e $m_T = 240$. Novamente, considerando valores de p variando entre $0,1$ e $0,9$, observamos que o valor de p que produziu as melhores estimativas para n_{AB} foi $0,64$, como no exemplo 7. Os resultados são apresentados na tabela 8, que contém os resumos da distribuição *a posteriori* de n_{AB}.

Tabela 8. Resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,50$, $\phi_D = 0,60$ e $p = 0,64$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	297,3879170	291	296	302	297	8,305818515	(281, 313)
1	1	1	1	297,4639802	291	296	302	297	8,298148624	(281, 313)
0	1	0	1	297,9256664	291	297	302	298	8,313878094	(281, 314)
1	0	1	0	296,8467474	290	296	301	296	8,297311539	(280, 312)
0,3	0,5	0,8	0,6	297,4048653	291	296	302	297	8,304903188	(281, 313)
0,8	3	2	0,4	297,4645787	291	296	302	297	8,296020421	(281, 313)
5	7	10	8	298,4565590	292	297	303	298	8,205871620	(282, 314)
15	18	20	31	303,5143909	297	302	308	303	8,098396479	(287, 319)
10	50	10	50	315,2744627	309	314	320	315	8,144472196	(299, 330)
50	10	50	10	283,8159872	278	283	288	283	7,437790983	(269, 298)
40	15	30	9	289,5969544	283	288	294	289	7,770982402	(274, 304)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão;

IC : Intervalo de credibilidade de 95%

Nesta tabela verificamos novamente que as estimativas *a posteriori* encontradas são bastante próximas do verdadeiro valor do parâmetro, além da influência dos valores dos hiperparâmetros α, β, α' e β'. Em particular, no caso em que β e β' assumem valores mais altos em relação a α e α', respectivamente, as estimativas ultrapassam o verdadeiro valor do parâmetro e em caso contrário observamos que as estimativas são inferiores.

Exemplo 9. Consideramos $\phi_C = 0,80$, $\phi_D = 0,85$ e as mesmas estatísticas obtidas no exemplo 6, ou seja, $m_C = 245, m_D = 262, m_{CD} = 215$ e $m_T = 292$. Neste caso nosso estudo evidenciou que o valor ideal para p foi $0,60$. Os resultados são apresentados na tabela 9, que contém os resumos *a posteriori* de n_{AB}.

Tabela 9. Resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,80$, $\phi_D = 0,85$ e $p = 0,60$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	297,4320115	295	296	298	297	2,581688906	(292, 302)
1	1	1	1	297,5429292	295	296	299	296	2,608059704	(292, 302)
0	1	0	1	297,58860115	295	296	298	297	2,620563916	(292, 302)
1	0	1	0	297,2776394	294	296	298	297	2,542813548	(292, 302)
0,3	0,5	0,8	0,6	297,4460726	295	296	298	297	2,584996870	(292, 302)
0,8	3	2	0,4	297,6303226	295	296	298	297	2,630562182	(292, 302)
5	7	10	8	299,0168827	296	298	300	298	2,931746951	(293, 304)
15	18	20	31	303,3676926	300	302	305	303	3,698967674	(296, 310)
10	50	10	50	311,5652776	307	310	314	311	4,745780075	(302, 320)
50	10	50	10	297,6834483	295	296	298	297	2,584713135	(292, 302)
40	15	30	9	298,6197915	296	297	299	298	2,812701395	(293, 304)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

Analisando os resultados obtidos na tabela acima, destacamos o caso em que $\alpha = \alpha' = 10$ e $\beta = \beta' = 50$. Além de produzir estimativas superiores em relação aos outros casos, fato já observado nas outras tabelas, o intervalo de credibilidade correspondente não contém o verdadeiro valor do parâmetro. Observamos ainda, pela tabela acima, que para os valores de ϕ_C e ϕ_D considerados neste caso, as amplitudes dos intervalos de credibilidade são menores em relação aos outros casos considerados nas tabelas anteriores.

3.1.3 Priori de Poisson

No caso da distribuição *a priori* de Poisson truncada em $\{M + 1, M + 2, ...\}$ para n_{AB} segue de (2.21) e (3.1) que, as funções $f(n_{AB})$, $f(n_{AB} + 1)$ e $Q(n_{AB})$, considerando as devidas simplificações, são dadas por

$$f(n_{AB}) = \frac{\lambda^{n_{AB}}}{(n_{AB} - m_T)!} \frac{\Gamma(n_{AB} - m_C + \beta)\Gamma(n_{AB} - m_D + \beta')}{\Gamma(n_{AB} + \alpha + \beta)\Gamma(n_{AB} + \alpha' + \beta')}, \tag{3.10}$$

$$f(n_{AB}+1) = \frac{\lambda^{n_{AB}+1}}{(n_{AB} - m_T + 1)!} \frac{\Gamma(n_{AB} - m_C + \beta + 1)\Gamma(n_{AB} - m_D + \beta' + 1)}{\Gamma(n_{AB} + \alpha + \beta + 1)\Gamma(n_{AB} + \alpha' + \beta' + 1)} \tag{3.11}$$

e

$$Q(n_{AB}) = \frac{\lambda(n_{AB} - m_C + \beta)(n_{AB} - m_D + \beta')}{(n_{AB} - m_T + 1)(n_{AB} + \alpha + \beta)(n_{AB} + \alpha' + \beta')}. \tag{3.12}$$

Com o objetivo de encontrar boas estimativas para o número de pares coincidentes, precisamos verificar inicialmente a influência do valor de λ na estimação do parâmetro. Dessa forma, observamos que para $\lambda = 301$ conseguimos obter, em todos os casos, estimativas próximas do verdadeiro valor do parâmetro n_{AB}. No próximo capítulo desenvolvemos um modelo bayesiano hierárquico que eliminará a escolha de λ.

Exemplo 10. Consideramos neste exemplo, $\phi_C = 0,10$, $\phi_D = 0,20$ e as estatísticas $m_C = 27, m_D = 54, m_{CD} = 6$ e $m_T = 75$ do exemplo 4. Os resumos da distribuição *a posteriori* de n_{AB} são apresentados na tabela 10.

Tabela 10. Resumos da distribuição *a posteriori de* n_{AB} para $\phi_C = 0,10$, $\phi_D = 0,20$ e $\lambda = 301$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	298,3623924	286	297	309	298	17,17803594	(264,332)
1	1	1	1	297,5365945	285	296	308	297	17,17088789	(263,331)
0	1	0	1	299,5087904	287	298	310	299	17,17478907	(265,333)
1	0	1	0	297,2080790	285	296	308	297	17,18158636	(263,330)
0,3	0,5	0,8	0,6	298,2860163	286	297	309	298	17,17709888	(264,331)
0,8	3	2	0,4	296,8133371	284	296	307	296	17,16785768	(263,330)
5	7	10	8	287,2007161	275	286	298	287	17,06093849	(253,320)
15	18	20	31	276,0102535	264	275	286	275	16,74744198	(243,308)
10	50	10	50	296,0967676	284	295	306	296	16,71581794	(263,328)
50	10	50	10	203,6255656	191	202	214	203	16,39521273	(171,235)
40	15	30	9	235,2631542	223	234	245	235	16,70864034	(202,268)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

Pela tabela acima verificamos que há influência dos valores atribuídos aos hiperparâmetros do modelo, de forma que melhores estimativas são determinadas quando consideramos valores menores para esses hiperparâmetros. Como já observados em todos os outros estudos feitos anteriormente, quanto maiores o valores de β e β' em relação aos valores de α e α', respectivamente, maiores as estimativas encontradas. Destacamos ainda as duas últimas linhas da tabela, nas quais os intervalos de credibilidade não contêm o verdadeiro valor do parâmetro.

Exemplo 11. Neste exemplo assumimos $\phi_C = 0,50$, $\phi_D = 0,60$ e as estatísticas $m_C = 143, m_D = 188, m_{CD} = 91$ e $m_T = 240$ do exemplo 5. A tabela 11 contém os resumos da distribuição *a posteriori* de n_{AB}.

Tabela 11. Resumos da distribuição *a posteriori de* n_{AB} para $\phi_C = 0,50$, $\phi_D = 0,60$ e $\lambda = 301$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	297,5369087	289	296	304	296	10,80314782	(277,319)
1	1	1	1	297,6627042	289	296	304	296	10,79440783	(277,319)
0	1	0	1	298,4455106	290	297	305	297	10,87853387	(277,320)
1	0	1	0	296,6309140	288	295	303	295	10,72670444	(276,318)
0,3	0,5	0,8	0,6	297,5612589	289	296	304	296	10,80253199	(277,319)
0,8	3	2	0,4	297,6626383	289	296	304	296	10,78962793	(277,319)
5	7	10	8	299,2903875	291	298	305	298	10,69307888	(279,320)
15	18	20	31	307,7801731	299	306	314	307	10,91989422	(287,329)
10	50	10	50	329,9398492	321	329	337	329	12,10591538	(306,354)
50	10	50	10	278,3289573	272	277	283	277	8,117906181	(263,294)
40	15	30	9	285,7960530	278	284	291	285	9,064292003	(268,304)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC :Intervalo de credibilidade de 95%.

Observando a tabela acima notamos novamente a influência dos valores atribuídos aos hiperparâmetros do modelo nas estimativas de n_{AB} e destacamos que, no caso em que $\alpha = \alpha' = 10$ e $\beta = \beta' = 50$ o intervalo de credibilidade não contém o verdadeiro valor do parâmetro.

Exemplo 12. Considerando $\phi_C = 0,80$, $\phi_D = 0,85$ e as estatísticas $m_C = 245, m_D = 262, m_{CD} = 215$ e $m_T = 292$ do exemplo 6, apresentamos na tabela 12 os resumos da distribuição *a posteriori* de n_{AB}.

Tabela 12. Resumos da distribuição *a posteriori de* n_{AB} para $\phi_C = 0,80$, $\phi_D = 0,85$ e $\lambda = 301$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	298,7956222	296	298	300	298	3,038149953	(293,305)
1	1	1	1	298,9423645	296	298	300	298	3,073116173	(293,305)
0	1	0	1	299,0055145	296	298	300	298	3,090842874	(293,305)
1	0	1	0	298,5894796	295	297	299	298	2,985767675	(293,304)
0,3	0,5	0,8	0,6	299,5894796	296	298	300	299	2,55473321	(294,305)
0,8	3	2	0,4	299,0610467	296	298	300	298	3,104057853	(293, 305)
5	7	10	8	300,9181950	297	300	302	300	3,511247288	(294, 308)
15	18	20	31	307,0692965	303	306	309	306	4,641185114	(298, 316)
10	50	10	50	320,2191077	315	319	323	319	6,526343490	(307, 333)
50	10	50	10	299,0443465	296	298	300	298	3,008379391	(293, 305)
40	15	30	9	300,3178313	297	299	301	299	3,324136968	(294, 307)

M:Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

Na tabela 12 notamos que para $\phi_C = 0,80$ e $\phi_D = 0,85$, as estimativas superam o valor verdadeiro do parâmetro na medida em que os valores dos hiperparâmetros vão aumentando e, em alguns casos, notamos que o intervalo de credibilidade não contém o verdadeiro valor do parâmetro.

3.2 Implementação via algoritmo *Gibbs sampling*

A título de ilustração, implementamos nesta seção o procedimento bayesiano via o método de simulação estocástica MCMC (Métodos de Monte Carlo via Cadeias de Markov), mais especificamente o algoritmo *Gibbs sampling.* O algoritmo *Gibbs sampling* foi introduzido por Geman e Geman (1984), como um algoritmo de simulação de distribuições multivariadas que aparecem em problemas de construção de imagens na Física Estatística. Por outro lado, Gelfand e Smith (1990) mostraram como o algoritmo pode

ser usado para simular de distribuições *a posteriori*. O algoritmo é baseado no seguinte resultado devido a Besag (1974). Suponhamos que $\underset{\sim}{\boldsymbol{\theta}} = (\theta_1, \theta_2, ..., \theta_k)$ seja o vetor de parâmetros de interesse e $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$ a distribuição *a posteriori*. Se $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$ for positiva em $\underset{\sim}{\Theta} = \underset{\sim}{\Theta}_1 \times \underset{\sim}{\Theta}_2 \times ... \times \underset{\sim}{\Theta}_k$, com $\underset{\sim}{\Theta}_i$ suporte da distribuição de θ_i para $i = 1, 2, ..., k$, então ela é unicamente determinada pelas distribuições condicionais completas $\pi(\theta_i|\theta_{(-i)}, D)$, onde $\theta_{(-i)}$ denota o vetor $(\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k)$ e D os dados.

O algoritmo de *Gibbs* se descreve do seguinte modo. Seja $\underset{\sim}{\boldsymbol{\theta}}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_k^{(0)})$ um valor inicial para o vetor $\underset{\sim}{\boldsymbol{\theta}}$. Procede-se iterativamente da seguinte forma:

(1) obtém-se $\theta_1^{(1)}$ de $\pi(\theta_1|\theta_2^{(0)}, ..., \theta_k^{(0)}, D)$,

obtém-se $\theta_2^{(1)}$ de $\pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, ..., \theta_k^{(0)}, D)$,

obtém-se $\theta_3^{(1)}$ de $\pi(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, ..., \theta_k^{(0)}, D)$,

...

obtém-se $\theta_k^{(1)}$ de $\pi(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_{k-1}^{(1)}, D)$.

Completa-se desse modo uma iteração do processo e uma transição de $\underset{\sim}{\boldsymbol{\theta}}^{(0)}$ para $\underset{\sim}{\boldsymbol{\theta}}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_k^{(1)})$.

(2) O processo anterior é repetido com $\underset{\sim}{\boldsymbol{\theta}}^{(1)}$ como vetor inicial, para obter um novo vetor $\underset{\sim}{\boldsymbol{\theta}}^{(2)} = (\theta_1^{(2)}, \theta_2^{(2)}, ..., \theta_k^{(2)})$ e assim haver uma transição de $\underset{\sim}{\boldsymbol{\theta}}^{(1)}$ para $\underset{\sim}{\boldsymbol{\theta}}^{(2)}$.

(3) Itera-se n vezes o ciclo de geração de observações aleatórias das distribuições condicionais, obtendo-se assim $\underset{\sim}{\boldsymbol{\theta}}^{(0)}, \underset{\sim}{\boldsymbol{\theta}}^{(1)}, \ldots, \underset{\sim}{\boldsymbol{\theta}}^{(n)}$.

A seqüência $\underset{\sim}{\boldsymbol{\theta}}^{(0)}, \underset{\sim}{\boldsymbol{\theta}}^{(1)}, \ldots, \underset{\sim}{\boldsymbol{\theta}}^{(n)}, \ldots$ é uma realização de uma cadeia de Markov com espaço de estados $\underset{\sim}{\Theta}$, função de transição

$$p(\underset{\sim}{\boldsymbol{\theta}}^{(n)}, \underset{\sim}{\boldsymbol{\theta}}^{(n+1)}) = \prod_{i=1}^{k} \pi(\theta_i^{(n+1)}|\theta_j^{(n)}, j > i, \theta_j^{(n+1)}, j < i, D)$$

e medida de probabilidade invariante ou medida de equilíbrio $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$.

A questão principal a ser analisada é como o algoritmo *Gibbs sampling* pode nos fornecer uma amostra aleatória da distribuição $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$? O que ocorre é que o método de amostragem *Gibbs* possibilita gerar realizações de uma cadeia de Markov de tal modo

que, à medida que o número de iterações aumenta, a cadeia se aproxima de sua condição de equilíbrio. Se numa determinada etapa a cadeia se encontra no estado de equilíbrio ou a convergência é atingida, então o vetor $\underset{\sim}{\boldsymbol{\theta}}^{(n)}$ gerado nessa etapa pode ser considerado como realização da distribuição $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$. Na realidade assume-se que a convergência é atingida numa iteração cuja distribuição esteja próxima da distribuição de equilíbrio e não no sentido formal da convergência. Contudo, sucessivas realizações de uma mesma cadeia ao longo do tempo (etapas) não constituem uma amostra aleatória da distribuição $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$. Com efeito, os vetores $\underset{\sim}{\boldsymbol{\theta}}^{(n)}$ que vão sendo gerados são correlacionados. Então, uma maneira de se obter uma amostra de $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$, é gerar duas, ou mais, cadeias, cada uma delas a partir de um estado inicial e utilizar como elementos amostrais algumas das observações resultantes dessas realizações, como descritas no que segue.

O período constituído pelas primeiras iterações até que a cadeia atinja o estado de equilíbrio é designado burn-in (período de aquecimento). O burn-in representa a quantidade de elementos gerados que será "descartada"durante o processo até a convergência, e seu valor depende da distribuição inicial.

Em geral, fazemos uma análise gráfica dos resultados da simulação para avaliar qual a quantidade ideal a ser descartada. Existem alguns procedimentos para se determinar o tamanho do "*burn in*", chamados diagnósticos de convergência. No estudo de simulação feito nesta dissertação, a convergência das cadeias foi verificada através do software CODA-Convergence Diagnostics and Output Analysis for Gibbs Sampling Output (Best, et al.,1995), utilizando o critério de convergência de Gelman Rubin que se encontra disponível no software CODA.

Para garantir uma independência aproximada entre os elementos gerados devemos considerar ainda, saltos entre esses elementos. Após burn-in e saltos, os elementos restantes podem ser considerados como uma amostra da distribuição *a posteriori* conjunta dos parâmetros, $\pi(\underset{\sim}{\boldsymbol{\theta}}|D)$.

Os programas utilizados para gerar as estimativas dos resumos *a posteriori* do parâmetro de interesse foram implementados via software R (versão 1.8.1) e estão disponíveis nos apêndices B.4 e B.5.

A título de ilustração e a fim de comparar as estimativas bayesianas de n_{AB} obtidas pelo *Gibbs sampling* e as obtidas pela relação recursiva, refizemos os exemplos de 4 a 12

utilizando agora o algoritmo *Gibbs sampling* e busca em tabela estática.

3.2.1 Priori Uniforme

Segue de (2.12), (2.13) e (2.14) que as distribuições condicionais de n_{AB}, ϕ_C e ϕ_D são tais que

- $\pi(n_{AB}|\phi_C,\phi_D,D) \propto \binom{n_{AB}}{m_T}[(1-\phi_C)(1-\phi_D)]^{n_{AB}}$; $m_T \leq n_{AB} \leq M$;
- ϕ_C, dados n_{AB},ϕ_D e D, tem distribuição Beta com parâmetros $m_C+\alpha$ e $n_{AB}-m_C+\beta$;
- ϕ_D dados n_{AB},ϕ_C e D, tem distribuição Beta com parâmetros $m_D+\alpha'$ e $n_{AB}-m_D+\beta'$.

A distribuição condicional de n_{AB}, dados ϕ_C,ϕ_D e D, não é conhecida, mas valores desta distribuição podem ser gerados pelo algoritmo busca em tabela estática.

Exemplo 13. Os dados e as estatísticas considerados neste exemplo são os mesmos do exemplo 4, isto é, dados $\phi_C = 0,10$, $\phi_D = 0,20$ as estatísticas obtidas por simulação foram $m_C = 27, m_D = 54, m_{CD} = 6$ e $m_T = 75$. Para estes valores geramos uma cadeia com 30.000 elementos, descartamos os 10.000 elementos iniciais e, considerando saltos de 10, obtivemos uma amostra final com 2.000 elementos.

As estimativas dos resumos da distribuição de probabilidades *a posteriori* de n_{AB} são dadas na tabela 13.

Tabela 13. Estimativas dos resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,10$ e $\phi_D = 0,20$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	260,168	201	247	308	233	76,849	(145, 436)
1	1	1	1	243,390	189	228	284	174	72,842	(138, 423)
0	1	0	1	287,467	226	278	345	203	78,937	(157, 446)
1	0	1	0	234,023	181	220	273	172	70,242	(134, 409)
0,3	0,5	0,8	0,6	256,312	199	241	305	217	77,430	(143, 432)
0,8	3	2	0,4	229,257	178	215	265	185	64,561	(133, 398)
5	7	10	8	154,562	133	149	168	144	30,610	(111, 219)
15	18	20	31	143,854	129	141	154	138	21,836	(111, 187)
10	50	10	50	261,983	226	256	292	223	49,858	(183, 380)
50	10	50	10	87,4862	84	87	90	85	4,7390	(80, 98)
40	15	30	9	95,353	89	94	98	93	6,795	(83, 109)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

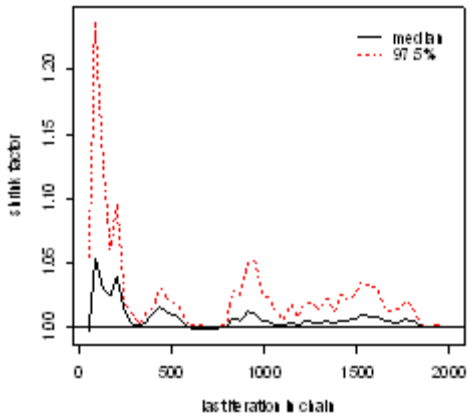

Figura 4. Gráfico de convergência de Gelman Rubin no caso $\alpha = \beta = \alpha' = \beta' = 1$.

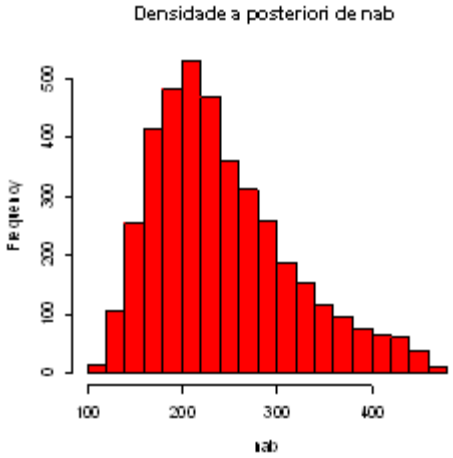

Figura 5. Histograma da estimativa da função de probabilidades a posteriori de n_{AB} para $\alpha = \beta = \alpha' = \beta' = 1$.

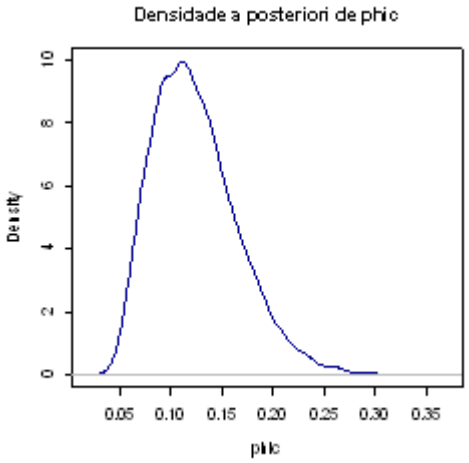

Figura 6. Gráfico da estimativa da função densidade a posteriori de ϕ_C no caso $\alpha = \beta = \alpha' = \beta' = 1$.

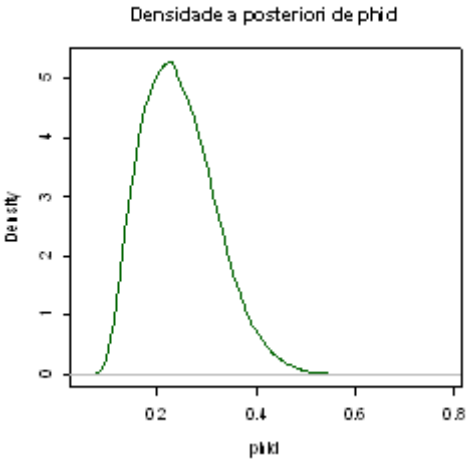

Figura 7. Gráfico da estimativa da função densidade a posteriori de ϕ_D no caso $\alpha = \beta = \alpha' = \beta' = 1$.

Comparando os dados da tabela 4 com os dados da tabela 13, concluímos que as estimativas bayesianas obtidas pelo método de relação recursiva são praticamente iguais às obtidas pelo algoritmo *Gibbs sampling.*

O gráfico da figura 4 indica que de fato houve a convergência dos elementos da cadeia gerados pelo algoritmo para $\alpha = \beta = \alpha' = \beta' = 1$, enquanto que as figuras 5, 6 e 7 representam, respectivamente, o comportamento das estimativas das distribuições *a posteriori* de n_{AB}, ϕ_C e ϕ_D para $\alpha = \beta = \alpha' = \beta' = 1$.

Para os demais valores de α, β, α' e β' da tabela 13 as estimativas das distribuições *a posteriori* de n_{AB}, ϕ_C e ϕ_D apresentaram comportamentos análogos ao caso em que $\alpha = \beta = \alpha' = \beta' = 1$.Utilizando os dados dos exemplos 5 e 6 constatamos, como no exemplo 13, que as metodologias relação recursiva e *Gibbs sampling* são equivalentes.

3.2.2 Priori Binomial

Segue de (2.17), (2.8) e (2.9) que as distribuições condicionais de n_{AB}, ϕ_C e ϕ_D são tais que

- $\pi(n_{AB}|\phi_C, \phi_D, D) \propto \binom{n_{AB}}{m_T}\binom{M}{n_{AB}}\left[\frac{p(1-\phi_C)(1-\phi_D)}{1-p}\right]^{n_{AB}}, \quad m_T \leq n_{AB} \leq M;$
- ϕ_C, dados n_{AB}, ϕ_D e D, tem distribuição Beta com parâmetros $m_C + \alpha$ e $n_{AB} - m_C + \beta$;
- ϕ_D dados n_{AB}, ϕ_C e D, tem distribuição Beta com parâmetros $m_D + \alpha'$ e $n_{AB} - m_D + \beta'$.

Exemplo 14. Neste exemplo consideramos os dados e as estatísticas do exemplo 4, isto é, consideramos $\phi_C = 0,10, \phi_D = 0,20$ e as estatísticas $m_C = 27, m_D = 54, m_{CD} = 6$ e $m_T = 75$. Variamos os valores de p e os valores entre $0,6$ e $0,7$ produziram , como no caso da relação recursiva, as melhores estimativas de n_{AB}. Para $p = 0,64$ geramos uma cadeia com 40.000 elementos, consideramos um burn in de 28.000 elementos e saltos de 6, obtendo uma amostra final de 2.000 elementos. Os resultados são apresentados na tabela 14.

Tabela 14. Estimativas dos resumos da distribuição *a posteriori* de n_{AB} para $\phi_C = 0,10, \phi_D = 0,20$ e $p = 0,64$.

α	β	α'	β'	M	$Q1$	Med	$Q3$	Mod	DP	IC
0,5	0,5	0,5	0,5	298,0200	291	298	305	296	10,4742	(278, 319)
1	1	1	1	297,4962	291	298	305	302	10,4278	(277,318)
0	1	0	1	298,0820	291	298	305	302	10,3663	(278, 319)
1	0	1	0	297,3782	290	298	305	295	10,4335	(277,318)
0,3	0,5	0,8	0,6	298,0100	291	298	305	295	10,3937	(278, 319)
0,8	3	2	0,4	297,1248	290	297	304	295	10,5157	(276, 318)
5	7	10	8	293,6035	286	294	301	291	10,5086	(273, 315)
15	18	20	31	289,4837	282	289	297	287	10,6182	(268, 311)
10	50	10	50	297,0883	290	297	304	288	10,2623	(277, 318)
50	10	50	10	262,2105	254	262	270	260	11,4557	(240, 285)
40	15	30	9	274,3735	267	274	282	272	10,9836	(252 , 296)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

Comparando os dados da tabela 7 com os dados da tabela 14 concluímos que o método da relação recursiva é equivalente ao *Gibbs sampling*. Utilizando os dados dos exemplos 5 e 6 constatamos, como no caso da priori uniforme para n_{AB}, que para $p = 0,60$ o *Gibbs sampling* e o método da relação recursiva são procedimentos equivalentes.

3.2.3 Priori de Poisson

Segue de (2.20), (2.8) e (2.9) que as distribuições condicionais de n_{AB}, ϕ_C e ϕ_D são tais que

- $\pi(n_{AB}|\phi_C, \phi_D, D) \propto \binom{n_{AB}}{m_T} \frac{\lambda^{n_{AB}}}{n_{AB}!}[(1-\phi_C)(1-\phi_D)]^{n_{AB}}$, $m_T \leq n_{AB} \leq M$;
- ϕ_C, dados n_{AB}, ϕ_D e D, tem distribuição Beta com parâmetros $m_C + \alpha$ e $n_{AB} - m_C + \beta$;

- ϕ_D dados n_{AB}, ϕ_C e D, tem distribuição Beta com parâmetros $m_D + \alpha'$ e $n_{AB} - m_D + \beta'$.

Utilizando os dados dos exemplos 4, 5 e 6 verificamos, como nos casos de *prioris* Uniforme e Binomial, que para $\lambda = 300$ o *Gibbs sampling* e o método de relação recursiva são equivalentes.

Capítulo 4

Estimação bayesiana hierárquica do número de indivíduos coincidentes: duas listas

Na seção 2.2 desenvolvemos um modelo bayesiano com o objetivo de estimar o número de indivíduos coincidentes em duas listas de elementos de uma população, n_{AB}. Para isso assumimos nas seções 2.2.1, 2.2.2 e 2.2.3 distribuições *a priori* Uniforme, Binomial e de Poisson, respectivamente, para n_{AB}, onde os hiperparâmetros do modelo são supostos conhecidos.

Como vimos no capítulo anterior, a escolha dos valores de tais hiperparâmetros deve ser cuidadosa para que possamos obter boas estimativas para o parâmetro de interesse n_{AB}. Para amenizar o problema da escolha desses hiperparâmetros vamos supor neste capítulo, que eles sejam variáveis aleatórias, assumindo determinadas distribuições de probabilidades. Desta maneira, desenvolvemos um modelo bayesiano hierárquico que possibilita a estimação de n_{AB}.

Nas seções abaixo vamos supor que os hiperparâmetros têm distribuição de probabilidades vagas ou não informativas.

4.1 *Priori* hierárquica Uniforme para o número de pares coincidentes

Para a construção do modelo hierárquico nesta seção, consideremos as seguintes suposições a *priori*.

1) Dados $\alpha, \beta, \alpha', \beta'$, $(\alpha > 0, \beta > 0, \alpha' > 0, \beta' > 0)$, n_{AB}, ϕ_C e ϕ_D são independentes, onde n_{AB} tem distribuição Uniforme no conjunto $\{0, 1, \ldots, M\}$, ϕ_C tem distribuição Beta com parâmetros α e β e ϕ_D tem distribuição Beta com parâmetros α' e β', isto é, dados $\alpha, \beta, \alpha', \beta'$, a distribuição a *priori* conjunta de n_{AB}, ϕ_C e ϕ_D é tal que

$$\pi(n_{AB}, \phi_C, \phi_D | \alpha, \beta, \alpha', \beta') \propto \left(\tfrac{1}{M+1}\right) \phi_C^{\alpha-1}(1-\phi_C)^{\beta-1}\phi_D^{\alpha'-1}(1-\phi_D)^{\beta'-1}$$

$$\propto \phi_C^{\alpha-1}(1-\phi_C)^{\beta-1}\phi_D^{\alpha'-1}(1-\phi_D)^{\beta'-1},$$

$0 < n_{AB} < M, 0 < \phi_C < 1, 0 < \phi_D < 1$;

2) α, β, α' e β' são independentes e identicamente distribuídos com distribuição Gama com parâmetros $\delta = 10^{-3}$ e $\Delta = 10^{-3}$, isto é, a distribuição conjunta de $\alpha, \beta, \alpha', \beta'$ é tal que

$$\pi(\alpha, \beta, \alpha', \beta') \propto \alpha^{10^{-3}-1}e^{-10^{-3}\alpha}\beta^{10^{-3}-1}e^{-10^{-3}\beta}(\alpha')^{10^{-3}-1}e^{-10^{-3}\alpha'}(\beta')^{10^{-3}-1}e^{-10^{-3}\beta'} =$$

$$= (\alpha\beta\alpha'\beta')^{10^{-3}-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')},$$

$\alpha > 0, \beta > 0, \alpha' > 0, \beta' > 0$.

Notamos que

$$E(\alpha) = E(\beta) = E(\alpha') = E(\beta') = \frac{\delta}{\Delta} = 1$$

e

$$Var(\alpha) = Var(\beta) = Var(\alpha') = Var(\beta') = \frac{\delta}{\Delta^2} = 10^3,$$

isto é, a distribuição adotada é não informativa. Logo, a distribuição a *priori* conjunta de $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta'$ é tal que

$$\pi(n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta') = \pi(n_{AB}, \phi_C, \phi_D|\alpha, \beta, \alpha', \beta')\pi(\alpha, \beta, \alpha', \beta')$$

$$\propto (\alpha\beta\alpha'\beta')^{10^{-3}-1}\phi_C^{\alpha-1}(1-\phi_C)^{\beta-1}\phi_D^{\alpha'-1}(1-\phi_D)^{\beta'-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')}, \qquad (4.1)$$

$0 \leq n_{AB} \leq M, 0 < \phi_C < 1, 0 < \phi_D < 1, \alpha > 0, \beta > 0, \alpha' > 0, \beta' > 0$. Das relações (2.1) e (4.1) segue que a distribuição a *posteriori* conjunta de $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha'$ e β' é tal que

$$\pi(n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta'|D) \propto L(n_{AB}, \phi_C, \phi_D|D)\pi(n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta')$$

$$\propto \tfrac{n_{AB}!}{(n_{AB}-m_T)!}(\alpha\beta\alpha'\beta')^{10^{-3}-1}\phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}\times$$

$$\times\phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')} \qquad (4.2)$$

$m_T \leq n_{AB} \leq M, 0 < \phi_C < 1, 0 < \phi_D < 1, \alpha > 0, \beta > 0, \alpha' > 0, \beta' > 0$.

De (4.2) segue que as distribuições condicionais são tais que

(i) $\pi(n_{AB}|\phi_C, \phi_D, \alpha, \beta, \alpha', \beta', D) \propto \frac{n_{AB}!}{(n_{AB}-m_T)!}[(1-\phi_C)(1-\phi_D)]^{n_{AB}}, \quad m_T \leq n_{AB} \leq M;$

(ii) $\pi(\phi_C|n_{AB}, \phi_D, \alpha, \beta, \alpha', \beta', D) \propto \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}, \quad 0 < \phi_C < 1$, isto é, dados $n_{AB}, \phi_D, \alpha, \beta, \alpha', \beta'$ e D, ϕ_C tem distribuição Beta com parâmetros $m_C + \alpha$ e $n_{AB} - m_C + \beta$;

(iii) $\pi(\phi_D|n_{AB}, \phi_C, \alpha, \beta, \alpha', \beta', D) \propto \phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}, \quad 0 < \phi_D < 1$,isto é, dados $n_{AB}, \phi_C, \alpha, \beta, \alpha', \beta'$ e D, ϕ_D tem distribuição Beta com parâmetros $m_D + \alpha'$ e $n_{AB} - m_D + \beta'$;

(iv) $\pi(\alpha|n_{AB}, \phi_C, \phi_D, \beta, \alpha', \beta', D) \propto \alpha^{10^{-3}-1}e^{-(10^{-3}-\ln\phi_C)\alpha}, \alpha > 0$, ou seja, dados $n_{AB}, \phi_C, \phi_D, \beta, \alpha', \beta'$ e D, α tem distribuição Gama com parâmetros 10^{-3} e $10^{-3} - \ln\phi_C$;

(v) $\pi(\beta|n_{AB},\phi_C,\phi_D,\alpha,\alpha',\beta',D) \propto \beta^{10^{-3}-1}e^{-(10^{-3}-\ln(1-\phi_C))\beta}$, $\beta>0$, isto é, dados n_{AB},ϕ_C, $\phi_D,\alpha,\alpha',\beta'$ e D, β tem distribuição Gama com parâmetros 10^{-3} e $10^{-3}-\ln(1-\phi_C)$;

(vi) $\pi(\alpha'|n_{AB},\phi_C,\phi_D,\alpha,\beta,\beta',D) \propto (\alpha')^{10^{-3}-1}e^{-(10^{-3}-\ln\phi_D)\alpha'}$, $\alpha'>0$, isto é, dados n_{AB},ϕ_C, $\phi_D,\alpha,\beta,\beta'$ e D, α' tem distribuição Gama com parâmetros 10^{-3} e $10^{-3}-\ln\phi_D$;

(vii) $\pi(\beta'|n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',D) \propto (\beta')^{10^{-3}-1}e^{-(10^{-3}-\ln(1-\phi_D))\beta'}$, $\beta'>0$, isto é, dados $n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha'$ e D, β' tem distribuição Gama com parâmetros 10^{-3} e $10^{-3}-\ln(1-\phi_D)$.

4.2 *Priori* hierárquica Binomial para o número de pares coincidentes

Nesta seção o modelo bayesiano hierárquico será construído do seguinte modo.

Suponhamos que a *priori*

1) Dados $\alpha,\beta,\alpha',\beta'$e p ($\alpha>0,\beta>0,\alpha'>0,\beta'>0$ e $0<p<1$), n_{AB},ϕ_C e ϕ_D são independentes, onde n_{AB} tem distribuição Binomial com parâmetros M e p, ϕ_C tem distribuição Beta com parâmetros α e β e ϕ_D tem distribuição Beta com parâmetros α' e β', isto é, dados $\alpha,\beta,\alpha',\beta'$ e p, a distribuição a *priori* conjunta de n_{AB},ϕ_C e ϕ_D é tal que

$$\pi(n_{AB},\phi_C,\phi_D|\alpha,\beta,\alpha',\beta',p) \propto \left(\begin{array}{c} M \\ n_{AB} \end{array}\right) p^{n_{AB}}(1-p)^{M-n_{AB}}\phi_C^{\alpha-1}(1-\phi_C)^{\beta-1}\phi_D^{\alpha'-1}(1-\phi_D)^{\beta'-1},$$

$n_{AB}=0,1,\ldots,M, 0<\phi_C<1, 0<\phi_D<1$;

2) $\alpha,\beta,\alpha',\beta'$ e p são independentes com $\alpha,\beta,\alpha',\beta'$ identicamente distribuídos com distribuição Gama com parâmetros $\delta=10^{-3}$ e $\Delta=10^{-3}$ e p com distribuição Uniforme no intervalo $(0,1)$, ou seja, a distribuição conjunta de $\alpha,\beta,\alpha',\beta'$ e p é tal que

$$\pi(\alpha,\beta,\alpha',\beta',p) \propto (\alpha\beta\alpha'\beta')^{10^{-3}-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')},$$

$\alpha>0,\beta>0,\alpha'>0,\beta'>0, 0<p<1$. Então, a distribuição *a priori* conjunta de $n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\beta'$ e p é tal que

$$\pi(n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\beta',p)=\pi(n_{AB},\phi_C,\phi_D|\alpha,\beta,\alpha',\beta',p)\pi(\alpha,\beta,\alpha',\beta',p)$$

$$\propto \begin{pmatrix} M \\ n_{AB} \end{pmatrix} (\alpha\beta\alpha'\beta')^{10^{-3}-1} p^{n_{AB}}(1-p)^{M-n_{AB}}\phi_C^{\alpha-1}\ (1-\phi_C)^{\beta-1}\ \times$$

$$\times\ \phi_D^{\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')}, \qquad (4.3)$$

$n_{AB}=0,1,\ldots,M, 0<\phi_C<1, 0<\phi_D<1, \alpha>0, \beta>0, \alpha'>0, \beta'>0, 0<p<1.$

Das relações (2.1) e (4.3) segue que a distribuição *a posteriori* conjunta de n_{AB}, ϕ_C , ϕ_D ,$\alpha, \beta, \alpha', \beta'$ e p é tal que

$$\pi(n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\beta',p|D)\propto L(n_{AB},\phi_C,\phi_D|D)\pi(n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\beta',p)$$

$$\propto \begin{pmatrix} n_{AB} \\ m_T \end{pmatrix}\begin{pmatrix} M \\ n_{AB} \end{pmatrix} (\alpha\beta\alpha'\beta')^{10^{-3}-1} p^{n_{AB}}(1-p)^{M-n_{AB}}\phi_C^{m_C+\alpha-1}\ \times$$

$$\times\ (1-\phi_C)^{n_{AB}-m_C+\beta-1}\phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')}, \qquad (4.4)$$

$m_T\leq n_{AB}\leq M, 0<\phi_C<1, 0<\phi_D<1, \alpha>0, \beta>0, \alpha'>0, \beta'>0, 0<p<1.$

Da relação(4.4) segue que as distribuições condicionais são tais que

(i) $\pi(n_{AB}|\phi_C,\phi_D,\alpha,\beta,\alpha',\beta',p,D)\propto \frac{1}{(n_{AB}-m_T)!(M-n_{AB})!}\left[\frac{p(1-\phi_C)(1-\phi_D)}{1-p}\right]^{n_{AB}}$, $m_T\leq n_{AB}\leq M$;

(ii) $\pi(\phi_C|n_{AB},\phi_D,\alpha,\beta,\alpha',\beta',p,D)\propto \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}$, $0<\phi_C<1$, isto é, a distribuição condicional de ϕ_C dados $n_{AB},\phi_D,\alpha,\beta,\alpha',\beta',p$ e D, é Beta com parâmetros $m_C+\alpha$ e $n_{AB}-m_C+\beta$;

(iii) $\pi(\phi_D|n_{AB},\phi_C,\alpha,\beta,\alpha',\beta',p,D)\propto \phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}$, $0<\phi_D<1$, ou

seja, a distribuição condicional de ϕ_D, dados $n_{AB}, \phi_C, \alpha, \beta, \alpha', \beta', p$ e D, é Beta com parâmetros $m_D + \alpha'$ e $n_{AB} - m_D + \beta'$;

(iv) $\pi(\alpha|n_{AB}, \phi_C, \phi_D, \beta, \alpha', \beta', p, D) \propto \alpha^{10^{-3}-1} e^{-(10^{-3} - \ln \phi_C)\alpha}$, $\alpha > 0$, ou seja, a distribuição condicional de α, dados $n_{AB}, \phi_C, \phi_D, \beta, \alpha', \beta', p$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln \phi_C$;

(v) $\pi(\beta|n_{AB}, \phi_C, \phi_D, \alpha, \alpha', \beta', p, D) \propto \beta^{10^{-3}-1} e^{-(10^{-3} - \ln(1-\phi_C))\beta}$, $\beta > 0$, isto é, a distribuição condicional de β, dados $n_{AB}, \phi_C, \phi_D, \alpha, \alpha', \beta', p$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln(1 - \phi_C)$;

(vi) $\pi(\alpha'|n_{AB}, \phi_C, \phi_D, \alpha, \beta, \beta', p, D) \propto (\alpha')^{10^{-3}-1} e^{-(10^{-3} - \ln \phi_D)\alpha'}$, $\alpha' > 0$, ou seja, a distribuição condicional de α', dados $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \beta', p$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln \phi_D$;

(vii) $\pi(\beta'|n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', p, D) \propto (\beta')^{10^{-3}-1} e^{-(10^{-3} - \ln(1-\phi_D))\beta'}$, $\beta' > 0$, ou seja, a distribuição condicional de β', dados $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', p$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln(1 - \phi_D)$;

(viii) $\pi(p|n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', D) \propto p^{n_{AB}}(1-p)^{M - n_{AB}}$, $0 < p < 1$, isto é, a distribuição condicional de p, dados $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta'$ e D, é Beta com parâmetros $n_{AB} + 1$ e $M - n_{AB} + 1$.

4.3 *Priori* hierárquica de Poisson para o número de pares coincidentes

Nesta seção o modelo bayesiano hierárquico será construído baseado nas seguintes suposições *a priori*

1) Dados $\alpha, \beta, \alpha', \beta'$ e λ ($\alpha > 0, \beta > 0, \alpha' > 0, \beta' > 0$ e $\lambda > 0$), n_{AB}, ϕ_C e ϕ_D são independentes, onde n_{AB} tem distribuição de Poisson truncada em $M+1, M+2, \ldots$, com parâmetro λ, ϕ_C tem distribuição Beta com parâmetros α e β e ϕ_D tem distribuição Beta com parâmetros α' e β'. Dessa forma, dados $\alpha, \beta, \alpha', \beta'$ e λ, a distribuição a *priori* conjunta de n_{AB}, ϕ_C e ϕ_D é tal que

$$\pi(n_{AB}, \phi_C, \phi_D|\alpha, \beta, \alpha', \beta', \lambda) \propto \frac{\lambda^{n_{AB}}}{n_{AB}!}\phi_C^{\alpha-1}\ (1-\phi_C)^{\beta-1}\ \phi_D^{\alpha'-1}\ (1-\phi_D)^{\beta'-1},$$

$n_{AB} = 0, 1, \ldots, M, 0 < \phi_C < 1, 0 < \phi_D < 1$;

2) $\alpha, \beta, \alpha', \beta'$ e λ são independentes com $\alpha, \beta, \alpha', \beta'$ identicamente distribuídos com distribuição Gama com parâmetros $\delta = 10^{-3}$ e $\Delta = 10^{-3}$ e λ com distribuição Gama com parâmetros $\delta' = 10^{-4}$ e $\Delta' = 10^{-4}$, ou seja, a distribuição conjunta de $\alpha, \beta, \alpha', \beta'$ e λ é tal que

$$\pi(\alpha, \beta, \alpha', \beta', \lambda) \propto (\alpha\beta\alpha'\beta')^{10^{-3}-1}\lambda^{10^{-4}-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')-10^{-4}\lambda}$$

$\alpha > 0, \beta > 0, \alpha' > 0, \beta' > 0, \lambda > 0$.

Observamos que considerando $\delta' = 10^{-4}$ e $\Delta' = 10^{-4}$ temos $E(\lambda) = 1$ e $Var(\lambda) = 10^4$, ou seja, a distribuição adotada para λ é não informativa. Assim, segue que a distribuição *a priori* conjunta de $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta'$ e λ é tal que

$$\pi(n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta', \lambda) = \pi(n_{AB}, \phi_C, \phi_D|\alpha, \beta, \alpha', \beta', \lambda)\pi(\alpha, \beta, \alpha', \beta', \lambda) \propto$$

$$\propto \frac{\lambda^{n_{AB}}}{n_{AB}!}(\alpha\beta\alpha'\beta')^{10^{-3}-1}\lambda^{10^{-4}-1}\phi_C^{\alpha-1}\ (1-\phi_C)^{\beta-1}\ \phi_D^{\alpha'-1}\ (1-\phi_D)^{\beta'-1}\times$$

$$\times\ e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')-10^{-4}\lambda}, \qquad (4.5)$$

$n_{AB} = 0, 1, \ldots, M$, $0 < \phi_C < 1$, $0 < \phi_D < 1$, $\alpha > 0$, $\beta > 0$, $\alpha' > 0$, $\beta' > 0$, $\lambda > 0$. Assim, segue de (2.1) e (4.5), que a distribuição *a posteriori* conjunta de $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha'$, β' e λ é tal que

$$\pi(n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\beta',\lambda|D) \propto L(n_{AB},\phi_C,\phi_D|D)\pi(n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\beta',\lambda)$$
$$\propto \frac{\lambda^{n_{AB}}}{n_{AB}!}(\alpha\beta\alpha'\beta')^{10^{-3}-1}\lambda^{10^{-4}-1}\phi_C^{m_C+\alpha-1}\ (1-\phi_C)^{n_{AB}-m_C+\beta-1}\ \times$$

$$\times\phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}e^{-10^{-3}(\alpha+\beta+\alpha'+\beta')-10^{-4}\lambda} \tag{4.6}$$

$0 \leq n_{AB} \leq M$, $0 < \phi_C < 1$, $0 < \phi_D < 1$, $\alpha > 0$, $\beta > 0$, $\alpha' > 0$, $\beta' > 0$, $\lambda > 0$.

De (4.6) segue que as distribuições condicionais para $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \alpha', \beta'$ e λ são, respectivamente

(i) $\pi(n_{AB}|\phi_C,\phi_D,\alpha,\beta,\alpha',\beta',\lambda,D) \propto \frac{1}{n_{AB}!}[\lambda(1-\phi_C)(1-\phi_D)]^{n_{AB}}, \quad m_T \leq n_{AB} \leq M$;

(ii) $\pi(\phi_C|n_{AB},\phi_D,\alpha,\beta,\alpha',\beta',\lambda,D) \propto \phi_C^{m_C+\alpha-1}(1-\phi_C)^{n_{AB}-m_C+\beta-1}, \quad 0 < \phi_C < 1$, isto é, a distribuição condicional de ϕ_C dados $n_{AB}, \phi_D, \alpha, \beta, \alpha', \beta', \lambda$ e D, é Beta com parâmetros $m_C + \alpha$ e $n_{AB} - m_C + \beta$;

(iii) $\pi(\phi_D|n_{AB},\phi_C,\alpha,\beta,\alpha',\beta',\lambda,D) \propto \phi_D^{m_D+\alpha'-1}(1-\phi_D)^{n_{AB}-m_D+\beta'-1}, \quad 0 < \phi_D < 1$, ou seja, a distribuição condicional de ϕ_D, dados $n_{AB}, \phi_C, \alpha, \beta, \alpha', \beta', \lambda$ e D, é Beta com parâmetros $m_D + \alpha'$ e $n_{AB} - m_D + \beta'$;

(iv) $\pi(\alpha|n_{AB},\phi_C,\phi_D,\beta,\alpha',\beta',\lambda,D) \propto \alpha^{10^{-3}-1}e^{-(10^{-3}-\ln\phi_C)\alpha}, \alpha > 0$, ou seja, a distribuição condicional de α, dados $n_{AB}, \phi_C, \phi_D, \beta, \alpha', \beta', \lambda$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln\phi_C$;

(v) $\pi(\beta|n_{AB},\phi_C,\phi_D,\alpha,\alpha',\beta',\lambda,D) \propto \beta^{10^{-3}-1}e^{-(10^{-3}-\ln(1-\phi_C))\beta}, \beta > 0$, isto é, a distribuição condicional de β, dados $n_{AB}, \phi_C, \phi_D, \alpha, \alpha', \beta', \lambda$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln(1-\phi_C)$;

(vi) $\pi(\alpha'|n_{AB},\phi_C,\phi_D,\alpha,\beta,\beta',\lambda,D) \propto (\alpha')^{10^{-3}-1}e^{-(10^{-3}-\ln\phi_D)\alpha'}, \alpha' > 0$, ou seja, a distribuição condicional de α', dados $n_{AB}, \phi_C, \phi_D, \alpha, \beta, \beta', \lambda$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln\phi_D$;

(vii) $\pi(\beta'|n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\lambda,D) \propto (\beta')^{10^{-3}-1}e^{-(10^{-3}-\ln(1-\phi_D))\beta'}, \beta' > 0$, ou seja, a distribuição condicional de β', dados $n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\lambda$ e D, é Gama com parâmetros 10^{-3} e $10^{-3} - \ln(1-\phi_D)$;

(viii) $\pi(\lambda|n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',D) \propto \lambda^{n_{AB}+10^{-4}-1}e^{-10^{-4}\lambda}, \lambda > 0$, isto é, a distribuição condicional de λ, dados $n_{AB},\phi_C,\phi_D,\alpha,\beta,\alpha',\beta'$ e D, é Gama com parâmetros $n_{AB} + 10^{-4}$ e 10^{-4}.

No próximo capítulo implementamos esse modelo.

Capítulo 5

Implementação do modelo bayesiano hierárquico

Neste capítulo apresentamos estimativas *a posteriori* do parâmetro de interesse, n_{AB}, considerando o modelo hierárquico desenvolvido no capítulo 4, para cada uma das *prioris* estudadas.

Os resultados apresentados nos exemplos abaixo foram obtidos através do algoritmo *Gibbs sampling*, descrito na seção 3.2.

Os dados utilizados são simulados e em todos os exemplos assumimos $n_A = 467, n_B = 611, n_{AB} = 297$. Em cada exemplo atribuímos às estatísticas m_C, m_D, m_{CD} e m_T os valores simulados no capítulo 3.

Os programas utilizados estão disponíveis nos apêndices C.1, C.2 e C.3.

5.1 *Priori* hierárquica uniforme para o número de pares coincidentes

Nesta seção apresentamos os resultados da implementação do modelo bayesiano hierárquico desenvolvido na seção 4.1.

Exemplo 15. Neste exemplo atribuímos a m_C, m_D, m_{CD} e m_T os valores atribuídos nos exemplos 4, 5 e 6, respectivamente. Em cada caso geramos uma cadeia com 30.000 elementos, descartamos os 10.000 elementos iniciais e, considerando saltos de 10, obtivemos

uma amostra de 2.000 elementos. As estimativas dos resumos da distribuição *a posteriori* de n_{AB} são dadas na tabela 15.

Tabela 15. Estimativas dos resumos da distribuição *a posteriori* de n_{AB}.

m_C	m_D	m_{CD}	m_T	M	$Q1$	Med	$Q3$	Mod	DP	IC
27	54	6	75	287,497	266	292	310	227	57,661	(163, 427)
143	188	91	240	296,8883	287	296	305	291	13,6096	(274, 326)
245	262	215	292	298,6623	296	298	300	298	5,0287	(284, 315)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

Comparando os resultados da primeira linha desta tabela com os da tabela 4 (exemplo 4), notamos que as estimativas fornecidas pelo modelo hierárquico são mais precisas do que as obtidas naquele exemplo. Por outro lado, observando os resultados da segunda e terceira linhas da tabela notamos que as estimativas obtidas são semelhantes às obtidas nos exemplos 5 e 6, respectivamente. Concluímos então que, neste caso, o modelo bayesiano hierárquico apresenta boas estimativas para o parâmetro de interesse e elimina a problemática da atribuição de valores para os hiperparâmetros.

5.2 *Priori* hierárquica Binomial para o número de pares coincidentes

Apresentamos, nesta seção, os resultados obtidos na implementação do modelo desenvolvido na seção 4.2.

Exemplo 16. Neste exemplo apresentamos as estimativas dos resumos da distribuição *a posteriori* de n_{AB}, considerando as mesmas estatísticas do exemplo anterior, ou seja, os valores das estatísticas utilizados nos exemplos 7, 8 e 9, respectivamente. Em cada caso geramos uma cadeia com 30.000 elementos, descartamos os 14.000 elementos iniciais e, considerando saltos de 8, obtivemos uma amostra de 2.000 elementos. Os resultados são apresentados na tabela 16.

Tabela 16. Estimativas dos resumos da distribuição *a posteriori* de n_{AB}.

m_C	m_D	m_{CD}	m_T	M	$Q1$	Med	$Q3$	Mod	DP	IC
27	54	6	75	298,0200	291	298	305	296	10,4742	(278, 319)
143	188	91	240	297,3608	292	297	303	297	8,1754	(281 , 313)
245	262	215	292	297,44975	296	297	299	297	2,55951	(293, 303)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

Comparando os resultados das linhas um, dois e três desta tabela aos resultados das tabelas 7 (exemplo 7), 8 (exemplo 8) e 9 (exemplo 9), respectivamente, notamos uma grande proximidade entre as estimativas obtidas. Isto significa que o modelo bayesiano hierárquico, além de eliminar a difícil escolha dos valores dos hiperparâmetros do modelo, apresenta boas estimativas para o parâmetro n_{AB}.

5.3 *Priori* hierárquica de Poisson para o número de pares coincidentes

O exemplo apresentado nesta seção contém os resumos *a posteriori* estimados através do modelo desenvolvido na seção 4.3.

Exemplo 17. Neste exemplo apresentamos as estimativas dos resumos da distribuição *a posteriori* de n_{AB}, considerando os mesmos dados dos exemplos anteriores, isto é, os dados utilizados nos exemplos 10, 11 e 12, respectivamente. Em cada caso geramos uma cadeia com 30.000 elementos, descartamos os 10.000 elementos iniciais e, considerando saltos de 10, obtivemos uma amostra de 2.000 elementos. A tabela 17 contém os resultados obtidos.

Tabela 17. Estimativas dos resumos da distribuição *a posteriori* de n_{AB}.

m_C	m_D	m_{CD}	m_T	M	$Q1$	Med	$Q3$	Mod	DP	IC
27	54	6	75	297,89457	286	297	309	298	17,25668	(264,332)
143	188	91	240	296,45887	289	296	304	296	10,97102	(277,319)
245	262	215	292	298,44658	296	298	300	298	3,056411	(293,305)

M :Média; Med :Mediana; Mod :Moda; DP :Desvio Padrão; IC : Intervalo de credibilidade de 95%.

Como podemos observar através da comparação das estimativas da primeira linha da tabela 17, com os resultados da tabela 10 (exemplo 10), as estimativas *a posteriori* obtidas nas duas tabelas são semelhantes, e muito próximas do verdadeiro valor do parâmetro n_{AB}. Da mesma forma, comparando as estimativas da segunda e terceira linhas da tabela 17 aos resultados das tabelas 11 (exemplo 11) e 12 (exemplo 12), respectivamente, notamos que o modelo hierárquico apresenta estimativas muito próximas às que já haviam sido obtidos nas tabelas 11 e 12. Neste caso, podemos concluir que o modelo hierárquico é adequado e evita a atribuição de valores para os hiperparâmetros.

Através dos exemplos deste capítulo, notamos que o modelo bayesiano hierárquico produz boas estimativas para o parâmetro de interesse, n_{AB}. Notamos ainda que os resultados obtidos neste capítulo são tão, ou mais, precisos daqueles obtidos no capítulo 3.

5.4 Considerações finais

Nosso objetivo nesta dissertação foi estimar o número de indivíduos coincidentes em duas amostras de uma população, que tratamos como listas. Para isso, desenvolvemos um modelo bayesiano considerando a possibilidade de que os dados referentes a cada indivíduo das listas pudessem ser preenchidos de maneira incorreta e descartando a chance de ocorrer uma falsa coincidência bareada nas informações dos indivíduos. Neste modelo, apresentamos estimativas do parâmetro de interesse, n_{AB}, considerando três diferentes

prioris para n_{AB}. Porém, a dificuldade de escolha dos valores para os hiperparâmetros de cada distribuição *a priori* nos fez optar pelo desenvolvimento de um modelo bayesiano hierárquico. Através das estimativas apresentadas por este último modelo, notamos claramente suas vantagens em relação ao modelo inicial. Além de facilitar a estimação do parâmetro de interesse, as estimativas obtidas estão muito próximas de seu verdadeiro valor e, em alguns casos, são mais precisas do que aquelas obtidas utilizando o primeiro modelo desenvolvido.

Sendo assim, concluímos que, no que se refere à estimação do número de indivíduos coincidentes em listas, o modelo bayesiano hierárquico apresenta-se como o mais adequado.

Comparando as resultados obtidos na implementação do modelo nos três diferentes casos de *prioris* estudadas, notamos que as estimativas obtidas nos três casos são muitos próximas. Isto nos leva a concluir que os três modelos considerados são equivalentes. Sendo assim, em uma situação prática, qualquer um dos três modelos escolhidos para a estimação do número de indivíduos concidentes em listas nos fornecerá bons resultados. Porém, observamos que o modelo que utiliza a distribuição *a priori* Uniforme apresenta um menor número de parâmetros e, portanto, sua implementação se torna mais fácil e rápida, podendo ser considerado preferencial entre todos.

Apêndice A

Programa utilizado no capítulo 2

A.1 - Programa para gerar as estatísticas e obter as estimativas de máxima verossimilhança dos parâmetros do modelo, utilizando a distribuição multinomial via software R.

```
set.seed(100)
# Dados
nab<-572;phic<-0.15;phid<-0.30
p1<-phic*(1-phid)
p2<-(1-phic)*phid
p3<-phic*phid
p4<-(1-phic)*(1-phid)
x<-numeric()
p.out<-p1+p2+p3+p4
mco<-mod<-mcd<-N0<-0
for (i in 1:nab)
{
x[i]<-runif(1)
if ((x[i] >= 0) && (x[i]<= p1))
{
mco<-mco+1
}
if ((x[i] > p1) && (x[i]<= (p1+p2)))
```

```
{
mod<-mod+1
}
if ((x[i] > (p1+p2)) && (x[i]<= (p1+p2+p3)))
{
mcd<-mcd+1
}
if ((x[i] > (p1+p2+p3)) && (x[i]<= 1))
{
N0<-N0+1
}
}
mc<-mco+mcd
md<-mod+mcd
mt<-mco+mod+mcd
nab<-mc+md-mt

mc; md; mt; mcd;

nabchapeu<-(mc*md)/mcd
phicchapeu<-mcd/md
phidchapeu<-mcd/mc

nabchapeu; phicchapeu; phidchapeu.
```

Apêndice B

Programas utilizados no capítulo 3

B.1 - Implementação do método de relação recursiva utilizando priori Uniforme para n_{AB} via software Maple.

```
restart;
f[nab]:=(nab!/(nab-mt)!)*(GAMMA(nab-mc+beta)*GAMMA(nab-md+betalinha))/
(GAMMA(nab+alpha+beta)*GAMMA(nab+alphalinha+betalinha));

Q[nab]:=simplify(eval(f[nab],nab=nab+1)/f[nab]);

alpha:=15; alphalinha:=20; beta:=18; betalinha:=31;
mc:=27; md:=54;
phic:=0.10; phid:=0.20;
M:=467;
distintos:=75;
n[distintos]:=distintos;
mt:=n[distintos];
f[mt]:=evalf(eval(f[nab],nab=mt));
for i from mt to M do f[i+1]:=evalf(eval(simplify(f[nab]*Q[nab]),nab=i)) od:
s:=sum('f[i]','i'=mt..M);
k:=1/s;        % constante normalizadora
for j from mt to M+1 do posteriori[j]:=k*f[j] od:
soma:=sum('posteriori[i]','i'=mt..M);

media:=sum('i*posteriori[i]','i'=mt..M);
```

variancia:=sum('((i-media)^2)*posteriori[i]','i'=mt..M):

desviopadrao:=sqrt(variancia);

% Construindo o gráfico de distribuição

for i from mt to M+1 do n[i]:=i od:

L1:=[n[mt],posteriori[mt]];

for i from n[mt+1] to n[M+1] do L1:=L1,[n[i],posteriori[i]] od:

L1:=[L1]:

plot(L1,nab=n[mt]..n[M],style=point,symbol=circle);

for i from n[mt] to n[M] do diferenca[i]:=posteriori[n[i+1]]-posteriori[n[i]] od;

for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.025) then inf:=j end if end do; inf;

for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.5) then mediana:=j end if end do; mediana;

for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.975) then sup:=j end if end do; sup;

for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.25) then q1:=j end if end do; q1;

for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.75) then q3:=j end if end do; q3;

B.2 - Implementação do método de relação recursiva utilizando priori Binomial para n_{AB} via software Maple.

restart;

f[nab]:=(((p^nab))*GAMMA(nab-mc+beta)*

GAMMA(nab-md+betalinha))/((nab-mt)!*(M-nab)!*

((1-p)^nab)*GAMMA(nab+alpha+beta)*

GAMMA(nab+alphalinha+betalinha));

Q[nab]:=simplify(eval(f[nab],nab=nab+1)/f[nab]);

alpha:=1; alphalinha:=1; beta:=1; betalinha:=1;

```
mc:=27; md:=54;
p:=0.5;
observados:=75;
n[observados]:=observados;
mt:=n[observados];
M:=467;
f[mt]:=evalf(eval(f[nab],nab=mt));
for i from mt to M do f[i+1]:=evalf(eval(simplify(f[nab]*Q[nab]),nab=i)) od:
s:=sum('f[i]','i'=mt..M);
k:=1/s;
for i from mt to M+1 do posteriori[i]:=k*f[i] od:
soma:=sum('posteriori[i]','i'=mt..M);

media:=sum('i*posteriori[i]','i'=mt..M);
variancia:=sum('((i-media)^2)*posteriori[i]','i'=mt..M):
desviopadrao:=sqrt(variancia);

for i from mt to M+1 do n[i]:=i od:
L1:=[n[mt],posteriori[mt]];
for i from n[mt+1] to n[M+1] do L1:=L1,[n[i],posteriori[i]] od:
L1:=[L1]:
plot(L1,nab=n[mt]..n[M],style=point,symbol=circle);
for i from n[mt] to n[M] do diferenca[i]:=posteriori[n[i+1]]-posteriori[n[i]] od:
for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.025) then inf:=j end if end
do; inf;
for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.5) then mediana:=j end if
end do;mediana;
for j from mt to M do if (sum('posteriori[i]','i'=mt..j)<=0.975) then sup:=j end if end
do;sup;
```

B.3 - Implementação do método de relação recursiva utilizando priori de Poisson para n_{AB} via software Maple.

O programa utilizado é análogo ao descrito no apêndice B, diferenciando-se apenas na função $f(n_{AB})$ que deve ser substituída pela função (3.12).

B.4 - Implementação do algoritmo Gibbs sampling considerando priori Uniforme para n_{AB} via software R.

O programa abaixo considera apenas uma cadeia para a geração dos dados.

```
set.seed(100)
bi<-20000 # burn-in
na<-40000 # tamanho da amostra a ser gerada
s<-10 # salto entre os valores amostrados para obter indep.
###Valores das estatísticas
alpha1<-1 # chute do valor de alpha
alpha2<-1 # chute do valor de alpha linha
beta1<-1 # chute do valor de beta
beta2<-1 # chute do valor de beta linha
n1<-467 #valor de na gerado da multinomial
n2<-611      #valor de nb gerado da multinomial
m<-min(n1,n2) #valor de M=min(na,nb)
ma<-23 # valor de mc
mb<-40 # valor de md
mab<-1 #valor de mcd
mT<-ma+mb-mab
c<-1
#valores iniciais
fi1o<-0.7
fi2o<-0.7
n12o<-400
#fi1o<-0.3
#fi2o<-0.3
#n12o<-576
```

```
fi1<-rep(0,na)
fi2<-rep(0,na)
n12<-rep(0,na)
fi1.out<-fi2.out<-n12.out<-numeric()
n12.test<-seq(mT,m,1)
lfx<-rep(0,length(n12.test))
fx<-rep(0,length(n12.test))
fx.out<-rep(0,na)
for (i in 1:length(n12.test))
{
    lfx[i]<-log(choose(n12.test[i],mT))+n12.test[i]*log(1-fi1o)+n12.test[i]*log(1-fi2o)
}
lfx<-lfx-max(lfx)
fx<-exp(lfx)
fx.out<-fx/sum(fx)
n12.trans<-rep(0,length(fx.out))
n12.trans[1]<-fx.out[1]
for (i in 2:length(fx.out))
{
    n12.trans[i]<-n12.trans[i-1]+fx.out[i]
}
n12.f<-0
prop<-0
uniforme<-runif(1)
if(uniforme < n12.trans[1])
    {
    prop<-n12.trans[1]
    n12.f<-n12.test[1]
    break
    }
for (j in 2:length(fx.out))
```

```
{
    if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
    {
    prop<-n12.trans[j]
    n12.f<-n12.test[j]
    break
    }
}
n12[1]<-n12.f
fi1[1]<-rbeta(1,ma+alpha1,n12[1]-ma+beta1)
fi2[1]<-rbeta(1,mb+alpha2,n12[1]-mb+beta2)
n12[1];fi1[1];fi2[1]

for     (i in 2:na)
{
    set.seed(i)
    lfx<-rep(0,length(n12.test))
    fx<-rep(0,length(n12.test))
    fx.out<-rep(0,na)
    for (w in 1:length(n12.test))
    {
    lfx[w]<-log(choose(n12.test[w],mT))+n12.test[w]*log(1-fi1[i-1])+
    n12.test[w]*log(1-fi2[i-1])
    }
    lfx<-lfx-max(lfx)
    fx<-exp(lfx)
    fx.out<-fx/sum(fx)
    n12.trans<-rep(0,length(fx.out))
    n12.trans[1]<-fx.out[1]
    for (k in 2:length(fx.out))
    {
```

```
        n12.trans[k]<-n12.trans[k-1]+fx.out[k]
        }
        n12.f<-0
        prop<-0
        uniforme<-runif(1,0,1)
        if(uniforme < n12.trans[1])
{
        prop<-n12.trans[1]
        n12.f<-n12.test[1]
        }
        for (j in 2:length(fx.out))
        {
        if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
        {
prop<-n12.trans[j]
n12.f<-n12.test[j]
break
}
}
n12[i]<-n12.f
        fi1[i]<-rbeta(1,ma+alpha1,n12[i]-ma+beta1)
        fi2[i]<-rbeta(1,mb+alpha2,n12[i]-mb+beta2)
}
for (k in 1:na)
{
if ((k > bi) && ((k-bi) %% s) == 0)
{
fi2.out<-rbind(fi2.out,fi2[k])
        fi1.out<-rbind(fi1.out,fi1[k])
        n12.out<-rbind(n12.out,n12[k])
        }
```

```
}
#n12.user<-mean(n12.out)
#n12.user<-median(n12.out)
#acf(n12.out) #autocorrelação de n12
#mean(n12.out)
#sd(n12.out)
#quantile(n12.out,c(0.025,0.25,0.50,0.75,0.975))
n121<-n12.out
fi11<-fi1.out
fi21<-fi2.out
hist(n12,main="Densidade a posteriori de nab ",col="red",xlab="nab")
#n122<-n12.out
#fi12<-fi1.out
#fi22<-fi2.out
### diagósticos de convergencia
library(coda)
a<-mcmc(n121)
b<-mcmc(n122)
ab<-mcmc.list(a,b)
summary(ab)
traceplot(ab)
gelman.diag(ab)
gelman.plot(ab)
autocorr(ab)
autocorr.plot(ab)
fi11m<-mcmc(fi11)
fi12m<-mcmc(fi12)
fi1<-mcmc.list(fi11m,fi12m)
summary(fi1)
traceplot(fi1)
gelman.diag(fi1)
```

```
gelman.plot(fi1)
autocorr(fi1)
fi21m<-mcmc(fi21)
fi22m<-mcmc(fi22)
fi2<-mcmc.list(fi21m,fi22m)
summary(fi2)
traceplot(fi2)
gelman.diag(fi2)
gelman.plot(fi2)
autocorr(fi2)
autocorr.plot(fi2)
```

B.5 - Implementação do algoritmo Gibbs sampling considerando priori Binomial para n_{AB} via software R.

O programa abaixo considera apenas uma cadeia para a geração dos dados.

```
set.seed(60)
bi<-28000        # burn-in
na<-40000        # tamanho da amostra a ser gerada
s<-6             # salto entre os valores amostrados para obter indep.
###Valores das estatísticas
alpha1<-1        # chute do valor de alpha
alpha2<-1        # chute do valor de alpha linha
beta1<-1         # chute do valor de beta
beta2<-1         # chute do valor de beta linha
n1<-467          # valor de na gerado da multinomial
n2<-611          # valor de nb gerado da multinomial
m<-min(n1,n2)        # valor de M=min(na,nb)
ma<-245          # valor de mc
mb<-262          # valor de md
mab<-215         # valor de mcd
```

```
mT<-ma+mb-mab
c<-1
p<-0.6
#valores iniciais
fi1o<-0.7
fi2o<-0.7
n12o<-400
fi1<-rep(0,na)
fi2<-rep(0,na)
n12<-rep(0,na)
fi1.out<-fi2.out<-n12.out<-numeric()
n12.test<-seq(mT,m,1) # seq de valores possíveis para n12, de mt ao min(n1,n2)
lfx<-rep(0,length(n12.test))
fx<-rep(0,length(n12.test))
fx.out<-rep(0,na)
for (i in 1:length(n12.test))
{
    lfx[i]<-log(choose(n12.test[i],mT))+log(choose(m,n12.test[i]))+n12.test[i]*
log(p)+n12.test[i]*log(1-fi1o)+n12.test[i]*log(1-fi2o)+(m-n12.test[i])*log(1-p)
}
lfx<-lfx-max(lfx)
fx<-exp(lfx)
fx.out<-fx/sum(fx)          # calcula a densidade em cada ponto
n12.trans<-rep(0,length(fx.out))
n12.trans[1]<-fx.out[1]          # assumindo um valor inicial para densidade de n12
for (i in 2:length(fx.out))
{
    n12.trans[i]<-n12.trans[i-1]+fx.out[i]
}
n12.f<-0
prop<-0
```

```
uniforme<-runif(1)
if(uniforme < n12.trans[1])
      {
            prop<-n12.trans[1]
      n12.f<-n12.test[1]
      break
      }
for (j in 2:length(fx.out))
{
      if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
      {
      prop<-n12.trans[j]
      n12.f<-n12.test[j]
      break
      }
}
n12[1]<-n12.f
fi1[1]<-rbeta(1,ma+alpha1,n12[1]-ma+beta1)
fi2[1]<-rbeta(1,mb+alpha2,n12[1]-mb+beta2)
n12[1];fi1[1];fi2[1]

for      (i in 2:na)
{
      lfx<-rep(0,length(n12.test))
      fx<-rep(0,length(n12.test))
      fx.out<-rep(0,na)
      for (w in 1:length(n12.test))
      {
lfx[w]<-log(choose(n12.test[w],mT))+log(choose(m,n12.test[w]))+n12.test[w]*
log(p)+n12.test[w]*log(1-fi1[i-1])+n12.test[w]*log(1-fi2[i-1])+(m-n12.test[w])*log(1-p)
      }
```

```
        lfx<-lfx-max(lfx)
        fx<-exp(lfx)
        fx.out<-fx/sum(fx)
        n12.trans<-rep(0,length(fx.out))
        n12.trans[1]<-fx.out[1]
        for (k in 2:length(fx.out))
        {
        n12.trans[k]<-n12.trans[k-1]+fx.out[k]
        }
        n12.f<-0
        prop<-0
        uniforme<-runif(1,0,1)
        if(uniforme < n12.trans[1])
        {
        prop<-n12.trans[1]
        n12.f<-n12.test[1]
        }
        for (j in 2:length(fx.out))
        {
        if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
        {
        prop<-n12.trans[j]
        n12.f<-n12.test[j]
        break
        }
        }
        n12[i]<-n12.f
        fi1[i]<-rbeta(1,ma+alpha1,n12[i]-ma+beta1)
        fi2[i]<-rbeta(1,mb+alpha2,n12[i]-mb+beta2)
    }
    for (k in 1:na)
```

```
{
    if ((k > bi) && ((k-bi) %% s) == 0)
    {
    fi2.out<-rbind(fi2.out,fi2[k])
    fi1.out<-rbind(fi1.out,fi1[k])
    n12.out<-rbind(n12.out,n12[k])
    }
}
n12.user<-mean(n12.out)
n12.user<-median(n12.out)
acf(n12.out) #autocorrelação de n12
mean(n12.out)
sd(n12.out)
quantile(n12.out,c(0.025,0.25,0.50,0.75,0.975))
n122<-n12.out
fi12<-fi1.out
fi22<-fi2.out
n12.out<-c(n121,n122) #moda
c<-max(n12.out)
d<-c-mT+1
freq<-rep(0,d)
for(j in 1:d) freq[n12.out[j]-mT+1]<-freq[n12.out[j]-mT+1]+1
m<-1
for(j in 2:d) if(freq[j]>freq[m]) m<-j
moda<-mT+m-1
moda
### diagósticos de convergencia
library(coda)
a<-mcmc(n121)
b<-mcmc(n122)
ab<-mcmc.list(a,b)
```

```
summary(ab)
traceplot(ab)
gelman.diag(ab)
gelman.plot(ab)
autocorr(ab)
autocorr.plot(ab)
acf(n121,n122)
fi11m<-mcmc(fi11)
fi12m<-mcmc(fi12)
fi1<-mcmc.list(fi11m,fi12m)
summary(fi1)
traceplot(fi1)
gelman.diag(fi1)
gelman.plot(fi1)
autocorr(fi1)
autocorr.plot(fi1)
fi21m<-mcmc(fi21)
fi22m<-mcmc(fi22)
fi2<-mcmc.list(fi21m,fi22m)
summary(fi2)
traceplot(fi2)
gelman.diag(fi2)
gelman.plot(fi2)
autocorr(fi2)
autocorr.plot(fi2)
hist(n12.out,main="Densidade a posteriori de nab ",col="red",xlab="nab")
plot(density(fi1),main="Densidade a posteriori de phic",col="darkblue",xlab="phic")
plot(density(fi2),main="Densidade a posteriori de phid",col="darkgreen",xlab="phid")
```

Apêndice C

Progamas utilizados no capítulo 5

C.1 - Implementação do modelo bayesiano hierárquico uniforme através do algoritmo Gibbs sampling via software R.

O programa fornecido abaixo apresenta uma única cadeia de valores porém, no momento da estimação mais de uma cadeia deve ser considerada.

```
set.seed(100)
bi<-10000        # burn-in
na<-30000        # tamanho da amostra a ser gerada
s<-10            # salto entre os valores amostrados para obter indep.
###Valores das estatísticas
n1<-467        # valor de n_A gerado da multinomial
n2<-611        # valor de n_B gerado da multinomial
m<-min(n1,n2)        # valor de M=min(na,nb)
ma<-27             # valor de mc
mb<-54             # valor de md
mab<-6             # valor de mcd
mT<-ma+mb-mab
c<-1
### Valores iniciais
fi1o<-0.3; fi2o<-0.6
```

```
n12o<-300
alpha1o<-0.5; alpha2o<-0.6
beta1o<-0.7; beta2o<-0.2
fi1<-rep(0,na)
fi2<-rep(0,na)
n12<-rep(0,na)
alpha1<-rep(0,na)
alpha2<-rep(0,na)
beta1<-rep(0,na)
beta2<-rep(0,na)
fi1.out<-fi2.out<-n12.out<-alpha1.out<-alpha2.out<-beta1.out<-beta2.out<-numeric()
n12.test<-seq(mT,m,1) # seq de valores possíveis para n12, de mt ao min(n1,n2)
lfx<-rep(0,length(n12.test))
fx<-rep(0,length(n12.test))
fx.out<-rep(0,na)
for (i in 1:length(n12.test))
{
    lfx[i]<-log(choose(n12.test[i],mT))+n12.test[i]*log(1-fi1[1])+n12.test[i]*log(1-fi2[1])
}
lfx<-lfx-max(lfx)
fx<-exp(lfx)
fx.out<-fx/sum(fx)
n12.trans[1]<-fx.out[1]
for (i in 2:length(fx.out))
{
    n12.trans[i]<-n12.trans[i-1]+fx.out[i]
}
n12.f<-0
prop<-0
uniforme<-runif(1)
if(uniforme < n12.trans[1])
```

```
        {
            prop<-n12.trans[1]
            n12.f<-n12.test[1]
            break
        }
    for (j in 2:length(fx.out))
    {
        if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
        {
            prop<-n12.trans[j]
            n12.f<-n12.test[j]
            break
        }
    }
    n12[1]<-n12.f
    fi1[1]<-rbeta(1,ma+alpha1o,n12[1]-ma+beta1o)
    fi2[1]<-rbeta(1,mb+alpha2o,n12[1]-mb+beta2o)
    n12[1];fi1[1];fi2[1];

    for      (i in 2:na)
    {
        set.seed(i)
        lfx<-rep(0,length(n12.test))
        fx<-rep(0,length(n12.test))
        fx.out<-rep(0,na)
        for (w in 1:length(n12.test))
        {

        lfx[w]<-log(choose(n12.test[w],mT))+n12.test[w]*log(1-fi1[i-1])+n12.test[w]*log(1-
fi2[i-1])
        }
```

```
lfx<-lfx-max(lfx)
fx<-exp(lfx)
fx.out<-fx/sum(fx)
n12.trans<-rep(0,length(fx.out))
n12.trans[1]<-fx.out[1]
for (k in 2:length(fx.out))
{
    n12.trans[k]<-n12.trans[k-1]+fx.out[k]
}
n12.f<-0
prop<-0
uniforme<-runif(1,0,1)
if(uniforme < n12.trans[1])
        {
            prop<-n12.trans[1]
            n12.f<-n12.test[1]

        }
for (j in 2:length(fx.out))
{
    if((n12.trans[j-1] <= uniforme) && (uniforme < n12.trans[j]))
    {
        prop<-n12.trans[j]
        n12.f<-n12.test[j]
        break
    }
}
n12[i]<-n12.f
fi1[i]<-rbeta(1,ma+alpha1[i-1],n12[i]-ma+beta1[i-1])
fi2[i]<-rbeta(1,mb+alpha2[i-1],n12[i]-mb+beta2[i-1])
alpha1[i]<-rgamma(1,10^(-3),10^(-3)-log(fi1))
```

```
    alpha2[i]<-rgamma(1,10^(-3),10^(-3)-log(fi2))
    beta1[i]<-rgamma(1,10^(-3),10^(-3)-log(1-fi1))
    beta2[i]<-rgamma(1,10^(-3),10^(-3)-log(1-fi2))
}
for (k in 1:na)
{
    if ((k > bi) && ((k-bi) %% s) == 0)
    {
        fi2.out<-rbind(fi2.out,fi2[k])
        fi1.out<-rbind(fi1.out,fi1[k])
        n12.out<-rbind(n12.out,n12[k])
        alpha1.out<-rbind(alpha1.out,alpha1[k])
        alpha2.out<-rbind(alpha2.out,alpha2[k])
        beta1.out<-rbind(beta1.out,beta1[k])
        beta2.out<-rbind(beta2.out,beta2[k])
    }
}
n12.user<-mean(n12.out)
n12.user<-median(n12.out)
acf(n12.out) #autocorrelação de n12
mean(n12.out)
sd(n12.out)
quantile(n12.out,c(0.025,0.25,0.50,0.75,0.975))
n121<-n12.out
fi11<-fi1.out
fi21<-fi2.out
alpha11<-alpha1.out
alpha21<-alpha2.out
beta11<-beta1.out
beta21<-beta2.out
# Moda
```

```
n12.out<-c(n121,n122)
c<-max(n12.out)
d<-c-mT+1
freq<-rep(0,d)
for(j in 1:d) freq[n12.out[j]-mT+1]<-freq[n12.out[j]-mT+1]+1
m<-1
for(j in 2:d) if(freq[j]>freq[m]) m<-j
moda<-mT+m-1
moda
### Diagósticos de convergência
library(coda)
a<-mcmc(n121)
b<-mcmc(n122)
ab<-mcmc.list(a,b)
summary(ab)
traceplot(ab)
gelman.diag(ab)
gelman.plot(ab)
autocorr(ab)
autocorr.plot(ab)
fi11m<-mcmc(fi11)
fi12m<-mcmc(fi12)
fi1<-mcmc.list(fi11m,fi12m)
summary(fi1)
traceplot(fi1)
gelman.diag(fi1)
gelman.plot(fi1)
autocorr(fi1)
autocorr.plot(fi1)
fi21m<-mcmc(fi21)
fi22m<-mcmc(fi22)
```

```
fi2<-mcmc.list(fi21m,fi22m)
summary(fi2)
traceplot(fi2)
gelman.diag(fi2)
gelman.plot(fi2)
autocorr(fi2)
autocorr.plot(fi2)
alpha11m<-mcmc(alpha11)
alpha12m<-mcmc(alpha12)
alpha1<-mcmc.list(alpha11m,alpha12m)
summary(alpha1)
traceplot(alpha1)
gelman.diag(alpha1)
gelman.plot(alpha1)
autocorr(alpha1)
autocorr.plot(alpha1)
alpha21m<-mcmc(alpha21)
alpha22m<-mcmc(alpha22)
alpha2<-mcmc.list(alpha21m,alpha22m)
summary(alpha2)
beta11m<-mcmc(beta11)
beta12m<-mcmc(beta12)
beta1<-mcmc.list(beta11m,beta12m)
summary(beta1)
```

C.2 - Implementação do modelo bayesiano hierárquico Binomial através do algoritmo Gibbs sampling via software R.

Este programa é análogo ao programa apresentado no apêndice G, devendo-se notar que as condicionais utilizadas aqui são as obtidas na seção 4.2. Por este motivo não o descrevemos neste apêndice.

C.3 - Implementação do modelo bayesiano hierárquico de Poisson através do algoritmo Gibbs sampling via software R.

Como no apêndice anterior, este programa é análogo ao programa apresentado no apêndice G, devendo-se notar que as condicionais utilizadas aqui são as obtidas na seção 4.3. Assim, omitimos sua descrição.

Referências Bibliográficas

[1] BELIN, T. R., RUBIN, D. B. (1995). A Method for Calibrating False-Match Rates in Record Linkage. *Journal of the American Statistical Association,* v. 90, n.**430**, p. 694-707.

[2] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Royal Statist. Soc. B*, **36**, 192 - 236.

[3] FELLEGI, I. P., SUNTER A. B. (1969). A Theory for Record Linkage. *American Statistical Association Journal,* p. 1183-1210.

[4] FIENBERG, S. E.; JOHNSON, M. S.; JUNKER, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. Journal Royal Statist. Soc. A, **162**, Part 3, p. 383-405.

[5] FORTINI, M.; LISEO, B.; NUCCITELLI, A.; SCANU, M.(2000). Modelling Issues in record linkage: A bayesian perspective. Relatório Técnico. Universidade de Roma "La Sapienza".

[6] FORTINI, M.; Liseo, B.; Nuccitelli, A.; Scanu, M.(2001). On Bayesian Record Linkage. *Research in Official Statistics,* p. 155-164.

[7] GELFAND, A. E.; SMITH, A. F. M. (1990). Sampling- based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398-409.

[8] GEMAN, S.; GEMAN, D. (1984). Stochastic relation, Gibbs distribution and the bayesian restoration of images. IEEE *Transcritions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

[9] JARO, M. A.(1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association,* v. 84, n.**406**, p. 414-420.

[10] LARSEN, M. D. (1999). Multiple imputation analysis of records linked using mixture models. Survey Methods Sect., Statistical Society of Canada, p. 65-71.

[11] LEE, A. J. (2002). Effect of list errors on the estimation of population size. *Biometrics*, **58**, p.185-191.

[12] LEE, A. J.; SEBER, G. A. F.; HOLDEN, J. K.; HUAKAU, J. T. (2001). Capture-Recapture, Epidemiology and List Mistaches: Several Lists. *Biometrics*, **57**, 707-713.

[13] MICHELETTI, L. R.(2003). Aplicação da metodologia da verossimilhança na prevalência do diabetes. Dissertação de Mestrado em Estatística, Centro de Ciências Exatas e Tecnologia, Departamento de Estatística, Universidade Federal de São Carlos.

[14] SCHEUREN. F.; ERNST; YOUNG, WINKLER, W. E. (1993). Regression Analysis of Data Files that are Computer Matched - Part I. *Survey Methodology,* **19**, 1, p. 39-58.

[15] SEBER, G. A. F.; HUAKAU, J. T.; SIMMONS, D. (2000). Capture-Recapture, Epidemiology, and List Mismatches: Two Lists. *Biometrics,* **57**, p. 1227-1232.

[16] SMITH, P.J.(1991). Bayesian analyses for a multiple capture-recapture model. *Biometrika*, **78**, 2, p. 399-407.

[17] TANCRED, A.; GUAGNANO, G.; LISEO, B. (2004). Inferenza statistica basata su dati prodotti mediante procedure di record linkage. Relatório Técnico. Universidade de Roma "La Sapienza".