Modelos HMM com dependência de segunda ordem: aplicação em genética

Daiane Aparecida Zuanetti

Orientador: Prof. Dr. Luís Aparecido Milan

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

fevereiro / 2006

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

Z293mh

Zuanetti, Daiane Aparecida.
Modelos HMM com dependência de segunda ordem: aplicação em genética / Daiane Aparecida Zuanetti. -- São Carlos : UFSCar, 2006.
90 p.

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

1. Estatística matemática. 2. Modelo markoviano oculto. 3. Redes probabilísticas. 4. Ordem de dependência. 5. Seleção de modelos. 6. MCMC. I. Título.

CDD: 519.5 (20ª)

Agradeço a Deus que me ofereceu condições emocionais, cognitivas, físicas, econômicas, culturais e sociais para completar, com êxito, todas as etapas da minha vida até a conclusão do mestrado e que me presenteou com uma família maravilhosa, que me ensinou, por gestos e palavras, o significado das palavras: confiança, respeito, humildade, ajuda, apoio, proteção, compreensão, incentivo, carinho, amor, responsabilidade e sinceridade. Aos meus pais, Luiz e Sebastiana, e irmãos, José Antonio e Patrícia, qualquer palavra de agradecimento seria insuficiente para expressar a importância de cada um em todos os momentos da minha vida. Aqui, basta registrar que eles são os exemplos de vida nos quais me espelho.

Agradeço também aos demais familiares e amigos com quem partilhei momentos muito felizes e divertidos e que me acompanharam em instantes de crescimento e superação. Aos professores e orientadores, agradeço por terem sido mediadores dos meus conhecimentos e do gosto pelo estudo. Agradeço também à FAPESP, Fundação de Amparo à Pesquisa do Estado de São Paulo, por ter financiado este projeto de mestrado.

Por fim, agradeço todas as pessoas que conheci durante minha caminhada e possibilitaram crecimento pessoal.

Resumo

A crescente necessidade do desenvolvimento de eficientes técnicas computacionais e estatísticas para analisar a profusão de dados biológicos transformaram o modelo Markoviano oculto (HMM), caso particular das redes bayesianas ou probabilísticas, em uma alternativa interessante para analisar seqüências de DNA. Uma razão do interesse no HMM é a sua flexibilidade em descrever segmentos heterogêneos da seqüência através de uma mesma estrutura de dependência entre as variáveis, supostamente conhecida. No entanto, na maioria dos problemas práticos, a estrutura de dependência não é conhecida e precisa ser também estimada. A maneira mais comum para estimação da estrutura de um HMM é o uso de métodos de seleção de modelos. Outra solução é a utilização de metodologias para estimação da estrutura de uma rede probabilística. Neste trabalho, propomos o HMM de segunda ordem e seus estimadores bayesianos, definimos o fator de Bayes e o DIC para seleção do HMM mais adequado a uma seqüência específica, verificamos seus desempenhos e a performance da metodologia proposta por Friedman e Koller (2003) em conjuntos de dados simulados e aplicamos estas metodologias em duas seqüências de DNA: o intron 7 do gene $\alpha - fetoprotein$ dos chimpanzés e o genoma do parasita *Bacteriophage lambda*, para o qual o modelo de segunda ordem é mais adequado.

Palavras-chave: Modelo Markoviano oculto; Redes probabilísticas; Ordem de dependência; Seleção de modelos; MCMC.

Abstract

An increasing need to develop efficient computational and statistical techniques and statistics to analyse the profusion biological data sets, transformed HMM into an interesting method to analyse DNA sequence. One reason for HMM interest is their flexibility in describing homogeneous segments inside the sequence through of a same dependence structure, supposely known. However, in many cases, the dependence structure is unknown and need to be estimated. This estimation can be made through model selection methods or techniques for probabilistic networks structure estimation. In this work, we propose the second order HMM and present the bayesian estimators to the involved parameters. We define the Bayes factor and DIC to select the most apropriate HMM to a specific sequence, verify the performance of these model selection methods and the dependence structure estimation method proposed by Friedman and Koller (2003), using simulated data sets and apply the techniques to intron 7 of the chimpanzee $\alpha - fetoprotein$ gene and genome of *Bacteriophage lambda*, to whichthe second-order HMM is more apropriated.

keywords: Hidden Markov model; Probabilistic networks; Order of dependence; Model selection; MCMC.

Sumário

Capítulo 1

Introdução

A metodologia dos modelos Markovianos ocultos (HMM, do inglês *Hidden Markov Models*) tem sido aplicada em uma grande variedade de problemas nas últimas duas décadas. Inicialmente aplicados em reconhecimento de fala (Rabiner, 1989) e econometria (Hamilton, 1989), os HMMs são, atualmente, muito utilizados na Biologia Molecular e Genética.

A vasta quantidade de seqüências de DNA disponível para análise, principalmente devido a projetos de seqüenciamento (entre eles o projeto Genoma Humano), e a crescente necessidade do desenvolvimento de técnicas computacionais eficientes e estatísticas para analisar esta profusão de dados biológicos transformam o HMM em uma alternativa interessante para analisar seqüências de DNA, que tem sido usado, sucessivamente, por Churchill (1989, 1992), que descreve a estrutura de uma seqüência de DNA através de um HMM; por Muri (1998), que mostra como este modelo é uma ferramenta útil para a segmentação de seqüências de DNA e por Boys *et al.* (2000, 2002), entre outros.

Uma razão do interesse no HMM é a sua flexibilidade em descrever segmentos heterogêneos da seqüência através de uma mesma estrutura, supostamente conhecida. Outra razão é o HMM ser um caso especial das redes probabilísticas ou bayesianas e, como tal, usualmente apresentado na forma de grafos que representam a dependência condicional das variáveis e fornecem uma descrição compacta e natural da dependência existente entre as variáveis aleatórias. Em particular, a estrutura de um modelo grafo possibilita um melhor entendimento das relações de independência condicional presentes no modelo associado.

O HMM é constituído por dois processos aleatórios, um observável e outro oculto, cujas variáveis se relacionam segundo uma estrutura de dependência específica. Se esta estrutura for conhecida, os algoritmos inferenciais de estimação dos parâmetros envolvidos nas distribuições de probabilidades condicionais e da seqüência de estados ocultos mais provável segundo os dados observados podem ser especificados. Boys *et al.* (2000), por exemplo, assumem que o processo observado se desenvolve como uma cadeia de Markov de primeira ordem condicionada à cadeia de Markov oculta, adotam a abordagem bayesiana e descrevem como informações *a priori* podem ser incorporadas no procedimento de estimação das probabilidades de transição entre os estados ocultos e observáveis e de identificação da seqüência de estados ocultos.

No entanto, na maioria dos problemas práticos, a estrutura de dependência não é conhecida e precisa ser também estimada. A maneira mais comum para estimação da estrutura de uma rede probabilística ou de um HMM é o uso de métodos de seleção de modelos. Contudo, estes métodos são eficientes somente quando o conjunto de prováveis e diferentes estruturas é pequeno. Quando há muitas possíveis estruturas de dependência para modelar um conjunto de dados específico, os métodos de seleção de modelos não são apropriados e uma metodologia mais geral e eficiente é necessária. Friedman e Koller (2003) propõem uma metodologia bayesiana para estimação da estrutura de uma rede probabilística baseada na probabilidade *a posteriori* de certas características da rede (presença de um arco entre duas variáveis, por exemplo, que indica dependência condicional entre elas).

Considerando, então, a importância e utilidade dos modelos Markovianos ocultos em várias áreas de pesquisa, o objetivo principal deste trabalho é apresentar e desenvolver os HMMs. Com este estudo, pretendemos, mais especificamente, introduzir os modelos com independência entre os estados observáveis e os modelos com dependência de Markov de primeira ordem, que são os modelos mais simples e, por isso, mais utilizados. Também é nosso objetivo propor modelos mais complexos com dependência de segunda ordem entre os estados observáveis ou entre os estados ocultos, apresentar estimadores para os parâmetros envolvidos nos modelos propostos e metodologias de estimação para modelos mais gerais e sofisticados e testá-los em conjuntos de dados simulados. Finalmente, desejamos verificar o desempenho de técnicas usuais de seleção de modelos, o fator de Bayes

e o DIC (do inglês, *deviance information criterion*), em comparar e selecionar, entre os modelos sugeridos, o modelo mais adequado aos dados.

Desta maneira, no Capítulo 2, definimos redes probabilísticas, introduzimos a metodologia proposta por Friedman e Koller (2003), aplicamo-la em um conjunto de variáveis binárias e apresentamos um estudo de simulação para avaliar a eficiência da metodologia proposta para estimação da estrutura de uma rede probabilística. No Capítulo 3, exibimos os HMMs como um caso particular das redes probabilísticas e definimos quatro diferentes modelos HMM e seus respectivos estimadores clássicos ou bayesianos. Dois métodos de seleção de modelos (fator de Bayes e DIC) para comparação e seleção do HMM mais adequado ao conjunto de dados observado são descritos no Capítulo 4. No Capítulo 5, aplicamos os modelos propostos em conjuntos de dados simulados e verificamos o desempenho dos métodos de seleção de modelos na comparação entre os modelos.

As aplicações a dados reais são apresentadas no Capítulo 6, no qual modelamos uma seqüência de DNA através de um HMM, descrevemos como definir as informações *a priori* dos parâmetros envolvidos no modelo e selecionamos, dentre um conjunto de HMMs pré-selecionados, o modelo mais adequado ao intron 7 do gene $\alpha - fetoprotein$ dos chimpanzés e ao genoma do parasita *Bacteriophage lambda*. As estimativas do modelo mais apropriado a cada seqüência também são apresentadas no Capítulo 6.

As considerações finais e as propostas de trabalhos futuros são apresentadas no Capítulo 7.

Capítulo 2

Redes Probabilísticas

Seja $\mathbf{X} = \{X_1, X_2, ..., X_p\}$ um conjunto de variáveis aleatórias. A este conjunto $\mathbf{X}$ podemos associar um grafo G definido como $G = (V, E)$, onde V denota o conjunto de vértices ou nós do grafo, tal que exista correspondência um a um entre os nós do grafo e as variáveis aleatórias, ou seja, $V = \{X_1, X_2, ..., X_p\}$ e E denota o conjunto de arestas, $\{e(i, l)\}$, onde i e l representam os nós X_i e X_l, $1 \leq i, l \leq p$.

Se as arestas são direcionadas são denominadas então de arcos, sendo que $e(i, l)$ significa que o arco é direcionado do nó i para o nó l, i é o pai do filho l. O conjunto de pais do filho l é denotado por $Pa_G(X_l)$. Um antecedente do nó i é um nó que tem como filho o nó i ou outro antecedente de i. Um descendente de i é um filho de i ou um filho de um descendente de i.

Dois nós, i e l, são adjacentes em G se E contém a aresta ou arco $e(i, l)$. Um caminho é uma seqüência de nós distintos $\{1, ..., m\}$ tal que exista um arco para cada par de nós subsequentes do caminho. Um caminho cujos nós inicial e final são os mesmos é denominado ciclo. Um ciclo direcionado é um ciclo de arcos cujas pontas têm a mesma direção. Se E contém somente arestas não direcionadas, o grafo G é um grafo não direcionado. Se, em contrapartida, E contiver somente arcos direcionados e nenhum ciclo direcionado, G é um grafo direcionado acíclico.

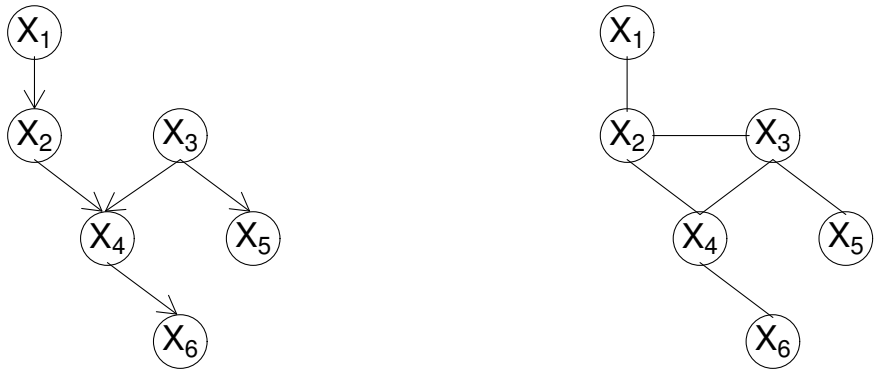
Neste capítulo introduzimos os modelos grafos direcionados acíclicos, conhecidos como Redes Probabilísticas ou Redes Bayesianas, discutimos a metodologia proposta por Friedman e Koller (2003) para estimação da estrutura de uma rede e apresentamos um conjunto de simulações que visam verificar a eficiência desta metodologia.

2.1 Redes probabilísticas

As redes probabilísticas ou redes bayesianas (Pearl, 1988) são uma representação através de um grafo da distribuição de probabilidades conjunta multivariada de $\mathbf{X}$ e fornecem uma descrição compacta e natural da dependência existente entre as variáveis aleatórias.

A rede probabilística ou bayesiana (tal nomenclatura não implicando, necessariamente, o uso de métodos bayesianos, mas o uso da fórmula de Bayes) para o conjunto de variáveis $\mathbf{X} = \{X_1, X_2, ..., X_p\}$ consiste de uma estrutura de rede e um conjunto de distribuições de probabilidades.

1. A estrutura de rede G é um grafo direcionado acíclico, no qual os nós correspondem às variáveis aleatórias, para $i = 1, ..., p$, e os arcos correspondem às dependências probabilísticas diretas entre as variáveis. Formalmente, a estrutura da rede representa o conjunto de especificações das relações de independência condicional para o modelo de probabilidades na forma de um grafo direcionado. Pode ser visto através do grafo, que uma variável X_i é condicionalmente independente de todas as outras variáveis, exceto das suas descendentes, dados os valores de seus pais.



(a) (b)

Figura 2.1: (a) A estrutura de G. (b) O grafo moral para G, onde os pais de X_4 foram "casados".

A estrutura G também pode ser representada pelo grafo moral, definido como um grafo não direcionado obtido de G através da inserção de arcos não direcionados

entre os pais não adjacentes de cada nó e desconsiderando a direção dos arcos direcionados restantes. O termo "moral" foi criado para denotar o "casamento dos pais não casados" (não adjacentes) do inglês "*marrying of unmarried parents*". Neste grafo, um nó é condicionalmente independente de todos os outros nós dada sua vizinhança.

A estrutura de uma rede probabilística e seu respectivo grafo moral são exibidos na Figura 2.1(a) e Figura 2.1(b), respectivamente. Na Figura 2.1(a) observamos que a variável X_4 é condicionalmente independente de X_1 e X_5 dados os valores de X_2 e X_3, ou seja, $X_4 \perp X_1|X_2$ e $X_4 \perp X_5|X_3$.

2. O conjunto P é constituído pelas distribuições de probabilidades *a priori* dos nós raiz (nós sem antecedente) e condicionais associadas a cada nó não raiz. Juntas, estas distribuições definem a distribuição de probabilidades conjunta de $\mathbf{X}$ que, dada a estrutura G e por uma propriedade da probabilidade condicional, é definida por

$$\Pr(\mathbf{x}) = \prod_{i=1}^{p} \Pr(x_i|x_1, ..., x_{i-1}), \tag{2.1}$$

onde $\mathbf{x} = \{x_1, x_2, ..., x_p\}$ é uma realização de $\mathbf{X} = \{X_1, X_2, ..., X_p\}$.

Como cada variável X_i é condicionalmente independente de todas as outras, exceto de suas variáveis descendentes, dados os valores de seus pais, relação que pode ser vista através do grafo, a equação (2.1) pode ser definida como

$$\Pr(\mathbf{x}) = \prod_{i=1}^{p} \Pr(x_i|\mathbf{pa}_G(X_i)), \tag{2.2}$$

onde $\mathbf{pa}_G(X_i)$ denota os valores assumidos pelos pais de X_i e os termos do produto (2.2) correspondem às distribuições condicionais ou *a priori*, se X_i for nó raiz, de P.

Consequentemente, o par (G, P) determina a distribuição conjunta de $\mathbf{X}$.

Se todas as variáveis são discretas, a distribuição de probabilidades condicional de cada nó pode ser representada em forma de tabela, que lista a probabilidade do nó filho assumir cada um dos diferentes valores de seu domínio para cada combinação dos valores de seus pais. Uma tabela para uma variável binária com k pais também binários contém 2^{k+1} probabilidades.

2.2 Estimação bayesiana da estrutura de uma rede probabilística

Um grande número de pesquisas dos últimos anos tem focalizado o problema de estimar a estrutura e os parâmetros a partir dos dados de uma rede (Buntine, 1996; Heckerman, 1998; entre outros). Estimar a estrutura é mais difícil do que estimar os parâmetros, assim como estimar ambos quando há dados perdidos ou variáveis ocultas é mais difícil do que estimá-los em situações nas quais todas as variáveis são observadas e os dados são completos.

A estrutura G de uma rede probabilística pode ser estimada através de métodos de seleção de modelos ou, quando desejamos calcular a probabilidade de uma característica específica da rede, por ponderação de modelos.

Em problemas com poucas variáveis e com uma quantidade significativa de dados, o uso da seleção de modelos produz resultados satisfatórios. No entanto, nos casos em que esta situação não ocorre, ou seja, a quantidade de dados é pequena em relação ao número de variáveis no modelo, há provavelmente muitos modelos que explicam os dados razoavelmente bem e a seleção de modelos realiza uma escolha arbitrária entre eles.

Dado que há muitas estruturas diferentes com probabilidades muito próximas, não podemos estimar uma única estrutura a partir dos dados. Além disso, em algumas situações, há muitas estruturas que são razoavelmente adequadas aos dados e enumerá-las é uma tarefa exaustiva. No entanto, podem haver algumas características da distribuição que são muito fortes e quase certas e, assim, já podemos fixá-las no modelo. Extrair estas características estruturais é o primeiro passo para a estimação da rede probabilística.

A probabilidade de uma característica estrutural, a presença de um arco por exemplo, dadas as observações D é calculdada por

$$\Pr(f|D) = \sum_{G} f(G)\pi(G|D), \tag{2.3}$$

onde G representa um modelo, $\pi(G|D)$ é a distribuição *a posteriori* de G e

$$f(G) = \begin{cases} 1 \text{ se a característica estiver em } G \\ 0 \text{ caso contrário} \end{cases}.$$

Se esta probabilidade é próxima de 1, então praticamente todos os modelos prováveis apresentam a característica. Caso contrário, se a probabilidade for baixa, sabemos que a característica está ausente na maioria dos modelos mais prováveis.

O número de estruturas para uma rede probabilística é exponencial ao número de variáveis aleatórias na rede. Assim, a soma em (2.3) pode ser calculada exatamente somente para redes com poucas variáveis. Alternativamente, esta soma pode ser aproximada considerando somente o subconjunto de estruturas mais prováveis. Uma aproximação deste subconjunto de estruturas mais prováveis pode ser obtida através do método Monte Carlo em Cadeias de Markov (MCMC), no qual definimos uma cadeia de Markov sobre o conjunto das estruturas cuja distribuição de equilíbrio é a distribuição *a posteriori* $\pi(G|D)$, geramos amostras desta cadeia e as usamos para estimar (2.3).

Apresentamos, a seguir, a metodologia proposta por Friedman e Koller (2003) para estimar a probabilidade *a posteriori* de certas características da estrutura da rede. A idéia básica do método é usar uma ordenação das variáveis da rede para separar o problema em dois problemas de mais fácil solução. Uma ordem $\prec$ corresponde a uma ordenação total das variáveis da rede, que coloca uma restrição na estrutura estimada de uma rede probabilística: se $X \prec Y$, nós restringimos a atenção às redes nas quais o arco entre X e Y, se existir, vai de X para Y. Assim, o problema de estimar a probabilidade *a posteriori* de uma característica da rede sobre o conjunto das estruturas pode ser dividido em dois subproblemas: estimar a probabilidade para as estruturas compatíveis com a dada ordem e somá-la considerando todas as possíveis ordens.

2.2.1 Estimação bayesiana da estrutura

Considere o problema de analisar a distribuição de um conjunto de variáveis $\mathbf{X} = (X_1, X_2, ..., X_p)$. Seja $D = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ um conjunto de dados completos, onde cada $\mathbf{x}_j$, para $j = 1, ..., n$, é uma realização completa das variáveis $X_1, X_2, ..., X_p$.

Uma rede probabilística $\mathcal{B}$ tem duas componentes: a estrutura G da rede e os valores dos parâmetros associados às distribuições de probabilidades condicionais, representados por $\boldsymbol{\theta}$. A parametrização $\boldsymbol{\theta}$ da rede varia . Em uma rede discreta de estrutura G, o vetor paramétrico $\boldsymbol{\theta}$ pode definir, por exemplo, uma distribuição multinomial $\boldsymbol{\theta}_{X_i|\mathbf{u}}$ para cada variável X_i e cada combinação dos valores $\mathbf{u}$ de $\mathbf{Pa}_G(X_i)$ (conjunto composto

pelas variáveis pais de X_i). Se considerarmos uma rede Gaussiana de domínio contínuo, na qual todas as variáveis aleatórias são normalmente distribuídas, X_i, para $i = 1, 2, ..., p$, será normalmente distribuída em torno de uma média que depende, linearmente, dos valores de seus pais e, assim, $\boldsymbol{\theta}_{X_i|\mathbf{u}}$ contém os coeficientes para uma combinação linear de $\mathbf{u}$ e um parâmetro de variância.

Para definirmos a distribuição *a priori* $\pi(\mathcal{B})$, precisamos definir uma distribuição *a priori* discreta para as estruturas G, $\pi(G)$, e para cada possível grafo G definir uma distribuição *a priori* sobre os possíveis valores de $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|G)$.

Distribuição *a priori* para a estrutura G

A distribuição *a priori* sobre as estruturas é, geralmente, considerada a menos importante das duas componentes de $\pi(\mathcal{B})$. A escolha mais simples e, por isso, mais comum é uma distribuição *a priori* uniforme sobre as estruturas (Heckerman, 1998). Para penalizar redes muito densas e complexas, podemos definir uma distribuição *a priori* usando uma probabilidade β da presença de cada arco, assim uma rede com h arcos tem distribuição *a priori* proporcional a $\beta^h(1-\beta)^{\binom{p}{2}-h}$ (Buntine, 1991). Uma distribuição *a priori* alternativa considera o número de opções na determinação das famílias de G, onde a família do nó X_i denota o conjunto de vértices que consiste de X_i e seus pais. Intuitivamente, se decidimos que o nó X_i tem k pais, há $\binom{p-1}{k}$ possíveis conjuntos de pais. Se escolhermos uniformemente entre estes conjuntos de pais, nós temos uma distribuição *a priori*

$$\pi(G) \propto \prod_{i=1}^{p} \binom{p-1}{|\mathbf{Pa}_G(X_i)|}^{-1}, \tag{2.4}$$

onde $|\mathbf{Pa}_G(X_i)|$ é o número de elementos no conjunto $\mathbf{Pa}_G(X_i)$.

Uma propriedade importante que estas distribuições *a priori* podem satisfazer é a modularidade estrutural.

- **Modularidade estrutural**

Se a distribuição *a priori* $\pi(G)$ satisfaz esta propriedade, então $\pi(G)$ pode ser escrita da forma

$$\pi(G) = \prod_i \rho_{X_i} \left(\mathbf{Pa}_G(X_i) \right), \tag{2.5}$$

onde $\rho_{X_i} \left(\mathbf{Pa}_G(X_i) \right)$ é uma distribuição sobre os possíveis conjuntos de pais de X_i, ou seja, a distribuição *a priori* é decomposta em um produto com um termo para cada variável do domínio da rede.

Distribuição *a priori* para o vetor paramétrico $\boldsymbol{\theta}$

Considere, agora, a distribuição *a priori* sobre os parâmetros, $\pi(\boldsymbol{\theta}|G)$. A forma destas distribuições *a priori* varia de acordo com o tipo de família paramétrica considerada. Em redes discretas com distribuição multinomial nos vértices, a suposição padrão é uma distribuição *a priori* Dirichlet sobre $\boldsymbol{\theta}_{X_i|\mathbf{u}}$ para cada variável X_i e cada combinação $\mathbf{u}$ de seus pais (Heckerman, 1998). Em redes gaussianas, podemos usar uma distribuição *a priori* Wishart (Heckerman e Geiger,1995). Nosso interesse é que a distribuição *a priori* satisfaça duas suposições básicas: independência global de parâmetros e modularidade dos parâmetros.

- **Independência global de parâmetros**

Sejam $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}$ os parâmetros que especificam o comportamento da variável X_i dadas as várias combinações de seus pais. Existe independência global de parâmetros se

$$\pi(\boldsymbol{\theta}|G) = \prod_i \pi \left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G \right). \tag{2.6}$$

Dada a independência dos parâmetros, construímos as distribuições *a priori* para os parâmetros de cada nó separadamente.

- **Modularidade dos parâmetros**

Sejam G e G' dois grafos nos quais $\mathbf{Pa}_G(X_i) = \mathbf{Pa}_{G'}(X_i) = \mathbf{U}$, então

$$\pi \left(\boldsymbol{\theta}_{X_i|\mathbf{U}}|G \right) = \pi \left(\boldsymbol{\theta}_{X_i|\mathbf{U}}|G' \right), \tag{2.7}$$

ou seja, a distribuição *a priori* dos parâmetros $\boldsymbol{\theta}_{X_i|\mathbf{U}}$ depende somente da família de X_i, composta por X_i e seus pais.

Distribuição *a posteriori* para a estrutura G

Definidas as distribuições *a priori* e usando a regra de Bayes, a distribuição *a posteriori* para a estrutura G pode ser escrita como

$$\pi\left(G|D\right) \propto \Pr\left(D|G\right)\pi(G), \tag{2.8}$$

onde o termo $\Pr\left(D|G\right)$ é a verossimilhança marginal dos dados dado G, e é definido como a integral da função de verossimilhança sobre todos os possíveis valores dos parâmetros para G,

$$\Pr\left(D|G\right) = \int \Pr\left(D|G,\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}|G\right)d\boldsymbol{\theta}. \tag{2.9}$$

O termo $\Pr\left(D|G,\boldsymbol{\theta}\right)$ é a probabilidade dos dados dada uma rede probabilística específica. Quando os dados são completos, este termo é um produto de probabilidades condicionais.

Usando as suposições anteriores, podemos mostrar (Heckerman *et al.*, 1995) que, se D é completo e $\pi\left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G\right)$ satisfaz independência dos parâmetros,

$$\Pr\left(D|G\right) = \prod_{i=1}^{p}\int\prod_{j=1}^{n}\Pr\left(x_{ij}|\mathbf{pa}_G(X_i)_j,\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}\right)\pi\left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G\right)d\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}. \tag{2.10}$$

Se a distribuição *a priori* $\pi\left(G\right)$ satisfaz modularidade estrutural, podemos concluir também que a probabilidade *a posteriori* se decompõe em

$$\pi\left(G|D\right) \propto \Pr\left(D|G\right)\pi(G) = \prod_{i=1}^{p} score\left(X_i,\mathbf{Pa}_G(X_i)|D\right), \tag{2.11}$$

onde

$$score\left(X_i,\mathbf{U}|D\right) = \rho_{X_i}\left(\mathbf{U}\right)\int\prod_{j=1}^{n}\Pr\left(x_{ij}|\mathbf{u}_j,\boldsymbol{\theta}_{X_i|\mathbf{U}}\right)\pi\left(\boldsymbol{\theta}_{X_i|\mathbf{U}}|G\right)d\boldsymbol{\theta}_{X_i|\mathbf{U}}, \tag{2.12}$$

$\mathbf{Pa}_G(X_i) = \mathbf{U}$ e $\mathbf{pa}_G(X_i) = \mathbf{u}$.

Para distribuições *a priori* tais como Dirichlet ou Wishart, o *score* $(X_i, \mathbf{U}|D)$ tem forma fechada e simples.

2.2.2 Ponderação bayesiana de modelos

O nosso principal objetivo é calcular a probabilidade *a posteriori* de alguma característica f sobre todos os possíveis grafos G. Esta probabilidade é

$$\Pr(f|D) = \sum_G f(G)\pi(G|D).$$

O problema é que o número de possíveis estruturas da rede probabilística é exponencial ao número de variáveis p. O número de estruturas pode diminuir se restringirmos atenção às estruturas G nas quais há um limite k no número de pais por nó. Seja, então, $\mathcal{G}_k$ o conjunto de todos os grafos com número de pais limitado por alguma constante k. O número de estruturas em $\mathcal{G}_k$ ainda é exponencial e a enumeração sobre o conjunto de possíveis estruturas é praticável somente para pequenos domínios.

Uma solução proposta por alguns pesquisadores (Madigan e Raftery, 1994; Madigan e York, 1995; Heckerman *et al.*, 1997) é aproximar esta exaustiva enumeração através da determinação de um conjunto $\mathcal{G}$ de estruturas mais prováveis e, então, estimar a massa relativa de cada estrutura em $\mathcal{G}$ que contém f através de

$$\Pr(f|D) \approx \frac{\sum_{G\in\mathcal{G}} f(G)\pi(G|D)}{\sum_{G\in\mathcal{G}} \pi(G|D)}. \tag{2.13}$$

Esta aproximação deixa aberta a questão sobre como construir $\mathcal{G}$. Madigan e York (1995) propuseram uma aproximação para $\mathcal{G}$ baseada no uso da simulação de Monte Carlo em Cadeias de Markov (MCMC). Green (1995) e Giudici *et al.* (2000) definiram uma cadeia de Markov do tipo *reversible jump* e estenderam a metodologia MCMC para casos nos quais a integração sobre os parâmetros é impraticável. Madigan *et al.* (1996) desenvolveu uma aproximação por amostragem MCMC sobre o espaço de grafos acíclicos parcialmente direcionados.

Estas construções de $\mathcal{G}$ por MCMC são uma aproximação que pode, a princípio, aproximar a verdadeira ponderação bayesiana de modelos através de amostragem da dis-

tribuição *a posteriori* da estrutura das redes. Elas têm sido usadas com sucesso em uma variedade de redes com pequenos domínios, geralmente 4 a 14 variáveis. No entanto, há alguns problemas que, potencialmente, limitam suas eficiências em domínios que envolvem muitas variáveis.

2.2.3 Distribuição *a posteriori* de uma característica consistente com uma ordem conhecida

Em vez de realizar ponderação de modelos sobre o espaço de todas estruturas, nós podemos considerar apenas as estruturas que são consistentes com alguma ordem total conhecida $\prec$ das variáveis, ou seja, consideramos estruturas G nas quais se $X_i \in \mathbf{Pa}_G(X_l)$, então, i precede l em $\prec$, ou $i \prec l$.

Primeiramente, consideramos o problema de calcular a probabilidade dos dados dada a ordem,

$$\begin{aligned} \Pr(D|\prec) &= \sum_{G\in\mathcal{G}_k} \Pr(D|G,\prec)\pi(G|\prec) \\ &= \sum_{G\in\mathcal{G}_k} \Pr(D|G)\pi(G|\prec). \end{aligned} \tag{2.14}$$

Embora esta soma seja restrita às redes com número de pais limitados e consistentes com $\prec$, o número de estruturas ainda é exponencialmente grande. Podemos, então, escolher uma estrutura G, consistente com $\prec$, através da escolha, independente, da família $\mathbf{U}$ para cada nó X_i.

Seja $\mathcal{G}_{k,\prec}$ o conjunto de estruturas em $\mathcal{G}_k$ consistentes com $\prec$. Usando (2.11) e (2.14) segue que

$$\Pr(D|\prec) = \sum_{G\in\mathcal{G}_{k,\prec}} \prod_{i=1}^{p} score\left(X_i, \mathbf{Pa}_G(X_i)|D\right). \tag{2.15}$$

A suposição de modularidade dos parâmetros em (2.7) estabelece que a escolha dos parâmetros para a distribuição de X_i depende apenas de sua família e não de toda a rede. Assim, se esta suposição for satisfeita, somar sobre todos os possíveis grafos

consistentes com $\prec$ é equivalente a somar sobre as possíveis escolhas da família de cada nó, cada uma com sua distribuição *a priori* dos parâmetros.

Dada a restrição sobre o tamanho da família, o conjunto composto por todos os possíveis conjuntos de pais para o nó X_i é

$$\mathcal{U}_{i,\prec} = \{\mathbf{U} : \mathbf{U} \prec X_i, |\mathbf{U}| \leq k\}, \tag{2.16}$$

onde $\mathbf{U} \prec X_i$ significa que todos os nós em $\mathbf{U}$ precedem X_i em $\prec$ e k é o número máximo de pais para cada variável. Assim, a probabilidade dos dados, dada uma ordem $\prec$, pode ser escrita como

$$\Pr(D|\prec) = \prod_{i=1}^{p} \sum_{\mathbf{U} \in \mathcal{U}_{i,\prec}} score(X_i, \mathbf{U}|D). \tag{2.17}$$

Intuitivamente, a igualdade estabelece que podemos somar sobre todas redes consistentes com $\prec$ através da soma sobre o conjunto de possíveis famílias para cada variável e, então, multiplicar este resultado para diferentes variáveis. Esta transformação nos permite calcular $\Pr(D|\prec)$ eficientemente. A expressão em (2.17) consiste de um produto com um termo para cada variável X_i, cada um destes é uma soma sobre todas as possíveis famílias para X_i. Dado o limite k sobre o número de pais, o número de possíveis famílias para a variável X_i é no máximo $\binom{p}{k} \leq p^k$. Portanto, a ordem de computações envolvidas em (2.17) é no máximo $p * p^k = p^{k+1}$. A computação de $\Pr(D|\prec)$ é um importante passo no algoritmo MCMC.

Probabilidade *a posteriori* de uma característica f

Para certos tipos de características f, podemos usar a técnica descrita na seção anterior para calcular a probabilidade de f estar presente na estrutura dados a ordem e os dados, $\Pr(f|\prec, D)$. Em geral, se $f(.)$ é uma característica, queremos calcular

$$\Pr(f|\prec, D) = \frac{\Pr(f, D|\prec)}{\Pr(D|\prec)}.$$

O cálculo de $\Pr(D|\prec)$ já foi discutido na seção anterior. O numerador de $\Pr(f|\prec, D)$ é uma soma sobre todas estruturas que contêm a característica e são consis-

tentes com a ordem,

$$\Pr(f, D|\prec) = \sum_{G \in \mathcal{G}_{k,\prec}} f(G)\pi(G|\prec)\Pr(D|G). \tag{2.18}$$

Este cálculo depende do tipo específico da característica f. A situação mais simples ocorre quando desejamos calcular a probabilidade *a posteriori* de uma particular escolha dos pais $\mathbf{U}$. Para isso precisamos somar sobre todos os grafos nos quais $\mathbf{Pa}_G(X_i) = \mathbf{U}$. Neste caso, podemos aplicar a mesma forma analítica fechada de (2.18). A única diferença é que restringimos $\mathcal{U}_{i,\prec}$ a $\{\mathbf{U}\}$. Assim,

$$\widehat{\Pr}(\mathbf{Pa}_G(X_i) = \mathbf{U}|D, \prec) = \frac{score(X_i, \mathbf{U}|D)}{\sum_{\mathbf{U}' \in \mathcal{U}_{i,\prec}} score(X_i, \mathbf{U}'|D)}. \tag{2.19}$$

Uma situação mais complexa ocorre quando desejamos calcular a probabilidade *a posteriori* de um arco $X_i \longrightarrow X_l$. Novamente podemos aplicar (2.18). A única diferença é que restringimos $\mathcal{U}_{j,\prec}$ de maneira que seja constituído somente pelos subconjuntos que contêm X_i. Então

$$\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec) = \frac{\sum_{\{\mathbf{U} \in \mathcal{U}_{l,\prec} : X_i \in \mathbf{U}\}} score(X_l, \mathbf{U}|D)}{\sum_{\mathbf{U} \in \mathcal{U}_{l,\prec}} score(X_l, \mathbf{U}|D)}. \tag{2.20}$$

2.2.4 Estimação da distribuição *a posteriori* de uma característica usando método MCMC

Introduzimos uma distribuição *a priori* uniforme sobre todas as possíveis ordens $\prec$ e definimos $\pi(G|\prec)$ como tendo a natureza das distribuições *a priori* usadas nas seções anteriores. Construímos, então, uma cadeia de Markov $\mathcal{M}$, com espaço de estados consistindo de todas as $p!$ ordens $\prec$, cuja distribuição de equilíbrio seja $\pi(\prec|D)$. Simulamos esta cadeia de Markov, obtendo uma seqüência de amostras $\prec_1, ..., \prec_T$ e, assim, podemos aproximar o valor esperado de qualquer função $g(\prec)$ como

$$E[g|D] \approx \frac{1}{T}\sum_{t=1}^{T} g(\prec_t).$$

Especificamente, podemos assumir $g(\prec) = \Pr(f|\prec, D)$ para alguma caracterís-

tica (arco) f e computar $g(\prec_t) = \Pr(f|\prec_t, D)$ como descrito nas seções anteriores.

Para construir a cadeia de Markov, usamos o algoritmo Metropolis-Hastings e definimos $q(\prec' \mid \prec)$ como a densidade geradora da transição de $\prec$ para $\prec'$. O algoritmo então aceita esta transição com probabilidade

$$\alpha(\prec, \prec') = \min\left[1, \frac{\pi(\prec' |D)\, q(\prec \mid \prec')}{\pi(\prec |D)\, q(\prec' \mid \prec)}\right],$$

e permanece em $\prec$ com probabilidade $1 - \alpha(\prec, \prec')$.

A densidade geradora da transição de $\prec$ para $\prec'$, $q(\prec' \mid \prec)$, pode ser construída de várias maneiras, baseada nos diferentes vizinhos no espaço de ordens. Em uma construção simples, consideramos somente operadores que invertem dois nós na ordem (mantendo todos os outros na mesma posição)

$$(i_1 ... i_l ... i_m ... i_p) \longmapsto (i_1 ... i_m ... i_l ... i_p), \prec \rightarrow \prec'.$$

Maiores detalhes sobre o algoritmo Metropolis-Hastings e sua construção pode ser encontrado em Chib e Greenberg (1995).

2.3 Estimação da estrutura de uma rede probabilística discreta

Considere o problema de analisar a distribuição de um conjunto de variáveis $\mathbf{X} = (X_1, X_2, ..., X_p)$, onde X_i, para $i = 1, ..., p$, tem distribuição de probabilidades *Bernoulli*(θ_{il}), com $\theta_{il} = P\left(X_i = 1|\mathbf{pa}_G^l(X_i), G\right)$, $\mathbf{pa}_G^l(X_i)$ denota a l-ésima combinação dos valores assumidos por $\mathbf{Pa}_G(X_i)$, para $i = 1, ..., p$ e $l = 1, ..., q_i$ e onde q_i é o número de possíveis combinações dos valores de $\mathbf{Pa}_G(X_i)$.

Seja $D = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ um conjunto completo de dados observados, onde cada $\mathbf{x}_j$, para $j = 1, ..., n$, é uma realização completa das variáveis $X_1, X_2, ..., X_p$. Suponha $\pi(\prec)$ uma distribuição *a priori* para $\prec$; $\pi(G| \prec)$ uma distribuição *a priori* para G consistente com $\prec$ e da natureza das distribuições *a priori* descritas na seção anterior e; $\pi(\boldsymbol{\theta}|G, \prec)$ uma distribuição *a priori* para $\boldsymbol{\theta}$, onde $\boldsymbol{\theta}$ é o vetor de parâmetros associados às distribuições de probabilidades condicionais, que satisfaz as condições de modularidade dos parâmetros

e de independência global de parâmetros.

A distribuição *a priori* conjunta para $(\prec, G, \boldsymbol{\theta})$ é dada por

$$\begin{aligned}
\pi(\ \prec\ , G, \boldsymbol{\theta}) &= \pi(\prec)\pi(G|\prec)\pi(\boldsymbol{\theta}|G, \prec) \\
&= \pi(\prec)\left(\prod_{i=1}^{p}\rho_{X_i}\left(\mathbf{Pa}_G(X_i)\right)\right)\left(\prod_{i=1}^{p}\pi\left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G, \prec\right)\right) \\
&= \pi(\prec)\left(\prod_{i=1}^{p}\rho_{X_i}\left(\mathbf{Pa}_G(X_i)\right)\pi\left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G, \prec\right)\right), \qquad (2.21)
\end{aligned}$$

onde $\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}$ é o vetor de probabilidades independentes de $X_i = 1$ dadas as várias combinações dos valores de seus pais $\mathbf{Pa}_G(X_i)$ e $\rho_{X_i}\left(\mathbf{Pa}_G(X_i)\right)$ é uma distribuição sobre os possíveis conjuntos de pais de X_i.

A distribuição *a posteriori* para $(\prec, G, \boldsymbol{\theta})$ é, então, dada por

$$\pi(\prec, G, \boldsymbol{\theta}|D) \propto \pi(\prec, G, \boldsymbol{\theta})\Pr(D|\prec, G, \boldsymbol{\theta}),$$

onde $\Pr(D|\prec, G, \boldsymbol{\theta})$ é a probabilidade dos dados dada uma rede probabilística específica. Como os dados são completos, ou seja, não temos dados perdidos e nenhuma variável oculta, este termo é o produto das probabilidades condicionais,

$$\Pr(D|\prec, G, \boldsymbol{\theta}) = \prod_{j=1}^{n}\prod_{i=1}^{p}\Pr(X_{ij} = x_{ij}|\mathbf{pa}_G(X_i)_j, \boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}), \qquad (2.22)$$

onde x_{ij} é o valor da variável X_i na j-ésima amostra.

Definidas $\Pr(D|\prec, G, \boldsymbol{\theta})$ e $\pi(\boldsymbol{\theta}|G, \prec)$, a distribuição *a posteriori* para $(\prec, G, \boldsymbol{\theta})$ pode ser escrita como

$$\begin{aligned}
\pi(\ \prec\ , G, \boldsymbol{\theta}|D) &\propto \pi(\prec)\pi(G|\prec)\pi(\boldsymbol{\theta}|G, \prec)\Pr(D|\prec, G, \boldsymbol{\theta}) \\
&= \pi(\prec)\left(\prod_{i=1}^{p}\rho_{X_i}\left(\mathbf{Pa}_G(X_i)\right)\pi\left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G\right)\right) \times \\
&\quad \times\left(\prod_{j=1}^{n}\prod_{i=1}^{p}\Pr(X_{ij} = x_{ij}|\mathbf{pa}_G(X_i)_j, \boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)})\right)
\end{aligned}$$

$$\pi(\prec, G, \boldsymbol{\theta}|D) \propto \pi(\prec) \prod_{i=1}^{p} \left(\begin{array}{c} \rho_{X_i}\left(\mathbf{Pa}_G(X_i)\right) \left(\prod_{j=1}^{n} \Pr(X_{ij} = x_{ij}|\mathbf{pa}_G(X_i)_j, \boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}) \right) \times \\ \times \pi\left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G, \prec\right) \end{array} \right). \tag{2.23}$$

Seja $\mathcal{G}_{k,\prec}$ o conjunto de estruturas em $\mathcal{G}_k$, conjunto de todos os grafos com número de pais limitado pela constante k e consistentes com $\prec$, a distribuição *a posteriori* marginal para $\prec$ pode ser obtida através da integração de $\pi(\prec, G, \boldsymbol{\theta}|D)$ em relação a $\boldsymbol{\theta}$ e da soma de $\pi(\prec, G, \boldsymbol{\theta}|D)$ para todas as estruturas $G \in \mathcal{G}_{k,\prec}$,

$$\begin{aligned} \pi(\ \prec\ |D) &= \sum_{G \in \mathcal{G}_{k,\prec}} \int \pi(\prec, G, \boldsymbol{\theta}|D)\ d\boldsymbol{\theta} \\ &\propto \pi(\prec) \sum_{G \in \mathcal{G}_{k,\prec}} \prod_{i=1}^{p} \left\{ \begin{array}{c} \rho_{X_i}\left(\mathbf{Pa}_G(X_i)\right) \times \\ \times \int \left[\begin{array}{c} \left(\prod_{j=1}^{n} \Pr(X_{ij} = x_{ij}|\mathbf{pa}_G(X_i)_j, \boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}) \right) \times \\ \times \pi\left(\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)}|G, \prec\right) d\boldsymbol{\theta}_{X_i|\mathbf{Pa}_G(X_i)} \end{array} \right] \end{array} \right\} \\ &= \pi(\prec) \sum_{G \in \mathcal{G}_{k,\prec}} \prod_{i=1}^{p} score(X_i, \mathbf{Pa}_G(X_i)|D) \end{aligned}$$

$$\pi(\prec |D) \propto \pi(\prec) \prod_{i=1}^{p} \sum_{\mathbf{U} \in \mathcal{U}_{i,\prec}} score\left(X_i, \mathbf{U}|D\right), \tag{2.24}$$

onde

$score\left(X_i, \mathbf{U}|D\right) = \rho_{X_i}\left(\mathbf{U}\right) \int \left(\prod_{j=1}^{n} \Pr(X_{ij} = x_{ij}|\mathbf{u}_j, \boldsymbol{\theta}_{X_i|\mathbf{U}}) \right) \pi\left(\boldsymbol{\theta}_{X_i|\mathbf{U}}|G, \prec\right) d\boldsymbol{\theta}_{X_i|\mathbf{U}}$ e

$\mathcal{U}_{i,\prec} = \{\mathbf{U} : \mathbf{U} \prec X_i, |\mathbf{U}| \leq k\}$ é o conjunto de todos os possíveis conjuntos de pais para X_i com no máximo k pais.

Se X_i, para $i = 1, ...p$, for a primeira variável em uma ordem específica de $\mathbf{X}$, o conjunto $\mathcal{U}_{i,\prec}$ se restringe ao conjunto $\emptyset$, $(\mathcal{U}_{i,\prec} = \emptyset)$, pois nenhuma variável antecede X_i na ordem e, consequentemente, nenhuma variável é candidata a ser pai de X_i. Já $\boldsymbol{\theta}_{X_i|\emptyset} = \theta_1 = \Pr\left(X_{ij} = 1\right)$, para $j = 1, ..., n$. Se X_i for a segunda variável da ordem, o

conjunto $\mathcal{U}_{i,\prec}$ é composto pelo conjunto $\emptyset$ e por $\left\{X_{(1)}\right\}$ $\left(\mathcal{U}_{i,\prec} = \left\{\emptyset; \left\{X_{(1)}\right\}\right\}\right)$, onde $X_{(1)}$ é a primeira variável da ordem. Neste caso, $\boldsymbol{\theta}_{X_i|\emptyset} = \theta_1 = \Pr\left(X_{ij} = 1\right)$, para $j = 1, ..., n$ e $\boldsymbol{\theta}_{X_i|\left\{X_{(1)}\right\}} = (\theta_1, \theta_2)$, onde $\theta_1 = \Pr\left(X_{ij} = 1|X_{(1)} = 0\right)$ e $\theta_2 = \Pr\left(X_{ij} = 1|X_{(1)} = 1\right)$, para $j = 1, ..., n$.

Caso X_i seja a terceira variável da ordem, $\mathcal{U}_{i,\prec} = \{\emptyset; \left\{X_{(1)}\right\}; \left\{X_{(2)}\right\}; \left\{X_{(1)}, X_{(2)}\right\}\}$, onde $X_{(1)}$ é a primeira variável da ordem e $X_{(2)}$ é a segunda variável da ordem. O número de elementos em $\boldsymbol{\theta}_{X_i|\mathbf{U}}$ associado a cada conjunto $\mathbf{U}$ de $\mathcal{U}_{i,\prec}$ é $2^{|\mathbf{U}|}$, onde $|\mathbf{U}|$ é o número de variáveis em $\mathbf{U}$, obrigatoriamente menor ou igual a k. Se X_i for a quarta variável na ordem, $\mathcal{U}_{i,\prec} = \{\emptyset; \left\{X_{(l)}\right\}$, para $l = 1, 2, 3; \left\{X_{(l)}, X_{(m)}\right\}$, para $l = 1, 2, 3$, $m = 1, 2, 3$ e $l \neq m; \left\{X_{(1)}, X_{(2)}, X_{(3)}\right\}\}$, onde $X_{(l)}$ é a l-ésima variável da ordem.

Se a posição de X_i na ordem for superior à quarta, o conjunto $\mathcal{U}_{i,\prec}$ é construído seguindo o mesmo procedimento descrito acima e sempre cuidando para que o número de variáveis em cada conjunto $\mathbf{U}$ de $\mathcal{U}_{i,\prec}$ não seja superior a k, número máximo de pais por variável.

A expressão matemática para $\Pr(X_{ij} = x_{ij}|\mathbf{u}_j, \boldsymbol{\theta}_{X_i|\mathbf{U}})$ varia para cada conjunto $\mathbf{U}$ de $\mathcal{U}_{i,\prec}$, assim como o número de elementos de $\boldsymbol{\theta}_{X_i|\mathbf{U}}$ depende do conjunto $\mathbf{U}$ considerado a cada instante. Como os elementos de cada $\boldsymbol{\theta}_{X_i|\mathbf{U}}$ são probabilidades, supostamente independentes, e como tal variam entre $(0, 1)$, é padrão assumir que a v-ésima componente das $2^{|\mathbf{U}|}$ probabilidades de $\boldsymbol{\theta}_{X_i|\mathbf{U}}$ tem distribuição Beta com parâmetros $\alpha_v > 0$ e $\beta_v > 0$. Desta maneira,

$$\pi\left(\boldsymbol{\theta}_{X_i|\mathbf{U}}|G\right) = \prod_{v=1}^{2^{|\mathbf{U}|}} \frac{1}{B\left(\alpha_v, \beta_v\right)} \theta_v^{\alpha_v - 1} \left(1 - \theta_v\right)^{\beta_v - 1} \tag{2.25}$$

e a função $score\left(X_i, \mathbf{U}|D\right)$ tem uma expressão matemática fechada para cada conjunto $\mathbf{U}$ de $\mathcal{U}_{i,\prec}$, para $i = 1, 2, ..., p$.

Considerando $\mathbf{U} = \emptyset$, a variável X_i, para $i = 1, 2, ..., p$, tem distribuição $Bernoulli\left(\theta_1\right)$, consequentemente $\boldsymbol{\theta}_{X_i|\mathbf{U}} = \{\theta_1\} \sim Beta\left(\alpha_1, \beta_1\right)$ e

$$\Pr(X_{ij} = x_{ij}|\boldsymbol{\theta}_{X_i|\mathbf{U}}) = \theta_1^{x_{ij}} \left(1 - \theta_1\right)^{1 - x_{ij}}. \tag{2.26}$$

Desta maneira,

$$
\begin{aligned}
score\left(X_i, \mathbf{U}|D\right) &= \rho_{X_i}\left(\mathbf{U}\right) \times \\
&\quad \times \int_0^1 \left(\prod_{i=1}^{p} \theta_1^{x_{ij}} \left(1-\theta_1\right)^{1-x_{ij}} \right) \frac{1}{B\left(\alpha_1, \beta_1\right)} \theta_1^{\alpha_1 - 1} \left(1-\theta_1\right)^{\beta_1 - 1} d\theta_1 \\
&= \rho_{X_i}\left(\mathbf{U}\right) \frac{1}{B\left(\alpha_1, \beta_1\right)} \int_0^1 \theta_1^{\sum_{j=1}^{n} x_{ij} + \alpha_1 - 1} \left(1-\theta_1\right)^{n - \sum_{j=1}^{n} x_{ij} + \beta_1 - 1} d\theta_1 \\
&= \rho_{X_i}\left(\mathbf{U}\right) \frac{B\left(n_{(1)} + \alpha_1, n_{(0)} + \beta_1\right)}{B\left(\alpha_1, \beta_1\right)}, \qquad (2.27)
\end{aligned}
$$

onde $n_{(a)}$ = número de casos em D tal que $X_i = a$, para $a = 0, 1$ e $B(\cdot, \cdot)$ é a função matemática *Beta*.

Considerando $\mathbf{U} = \{X_l\}$, tal que X_l precede a variável X_i na ordem $\prec$ e $l \neq i$,

$$
X_i \sim \begin{cases} Bernoulli(\theta_1), \text{ se } X_l = 0 \\ Bernoulli(\theta_2), \text{ se } X_l = 1 \end{cases}.
$$

Consequentemente, $\boldsymbol{\theta}_{X_i|\mathbf{U}} = \{\theta_1, \theta_2\}$, $\theta_v \sim Beta\left(\alpha_v, \beta_v\right)$, para $v = 1, 2$ e

$$
\Pr(X_{ij} = x_{ij} | X_{lj} = x_{lj}, \boldsymbol{\theta}_{X_i|\mathbf{U}}) = \theta_1^{x_{ij}\left(1-x_{lj}\right)} \left(1-\theta_1\right)^{(1-x_{ij})\left(1-x_{lj}\right)} \theta_2^{x_{ij}x_{lj}} \left(1-\theta_2\right)^{(1-x_{ij})x_{lj}}. \qquad (2.28)
$$

Assim,

$$
\begin{aligned}
score\left(X_i, \mathbf{U}|D\right) &= \rho_{X_i}\left(\mathbf{U}\right) \times \\
&\quad \times \int_0^1 \int_0^1 \left\{ \begin{array}{c} \left(\prod_{v=1}^{2} \frac{1}{B(\alpha_v, \beta_v)} \theta_v^{\alpha_v - 1} \left(1-\theta_v\right)^{\beta_v - 1} \right) \times \\ \times \prod_{j=1}^{n} \left(\begin{array}{c} \theta_1^{x_{ij}\left(1-x_{lj}\right)} \left(1-\theta_1\right)^{(1-x_{ij})\left(1-x_{lj}\right)} \times \\ \theta_2^{x_{ij}x_{lj}} \left(1-\theta_2\right)^{(1-x_{ij})x_{lj}} \end{array} \right) \end{array} \right\} d\theta_1 d\theta_2
\end{aligned}
$$

$$
\begin{aligned}
score\left(X_i, \mathbf{U}|D\right) &= \frac{\rho_{X_i}\left(\mathbf{U}\right)}{\prod_{v=1}^{2} B\left(\alpha_v, \beta_v\right)} \int_0^1 \left\{ \theta_2^{\sum_{j=1}^{n} x_{ij}x_{lj}+\alpha_2-1} \left(1-\theta_2\right)^{\sum_{j=1}^{n} x_{lj} - \sum_{j=1}^{n} x_{ij}x_{lj}+\beta_2-1} d\theta_2 \right\} \times \\
&\quad \times \int_0^1 \left\{ \theta_1^{\sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} x_{ij}x_{lj}+\alpha_1-1} \left(1-\theta_1\right)^{n-\sum_{j=1}^{n} x_{lj} - \sum_{j=1}^{n} x_{ij} + \sum_{j=1}^{n} x_{ij}x_{lj}+\beta_1-1} d\theta_1 \right\} \\
&= \rho_{X_i}\left(\mathbf{U}\right) \frac{B\left(n_{(1,0)}+\alpha_1, n_{(0,0)}+\beta_1\right) B\left(n_{(1,1)}+\alpha_2, n_{(0,1)}+\beta_2\right)}{\prod_{v=1}^{2} B\left(\alpha_v, \beta_v\right)}, \qquad (2.29)
\end{aligned}
$$

onde $n_{(a,b)}$ = número de casos em D tal que $X_i = a$ e $X_l = b$, para $a = 0, 1$, $b = 0, 1$.

Para $\mathbf{U} = \{X_l, X_m\}$, tal que X_l e X_m precedem X_i na ordem $\prec$ e $l \neq m \neq i$,

$$
X_i \sim \begin{cases} Bernoulli(\theta_1), \text{ se } X_l = X_m = 0 \\ Bernoulli(\theta_2), \text{ se } X_l = 1 \text{ e } X_m = 0 \\ Bernoulli(\theta_3), \text{ se } X_l = 0 \text{ e } X_m = 1 \\ Bernoulli(\theta_4), \text{ se } X_l = X_m = 1 \end{cases}.
$$

Consequentemente, $\boldsymbol{\theta}_{X_i|\mathbf{U}} = \{\theta_1, \theta_2, \theta_3, \theta_4\}$, $\theta_v \sim Beta\left(\alpha_v, \beta_v\right)$, para $v = 1, 2, 3, 4$ e

$$
\begin{aligned}
\Pr(X_{ij} &= x_{ij}|X_{lj} = x_{lj}, X_{mj} = x_{mj}, \boldsymbol{\theta}_{X_i|\mathbf{U}}) = \theta_1^{x_{ij}\left(1-x_{lj}\right)(1-x_{mj})} \times \\
&\times \left(1-\theta_1\right)^{(1-x_{ij})\left(1-x_{lj}\right)(1-x_{mj})} \theta_2^{x_{ij}x_{lj}(1-x_{mj})} \left(1-\theta_2\right)^{(1-x_{ij})x_{lj}(1-x_{mj})} \\
&\times \theta_3^{x_{ij}\left(1-x_{lj}\right)x_{mj}} \left(1-\theta_3\right)^{(1-x_{ij})\left(1-x_{lj}\right)x_{mj}} \times \\
&\times \theta_4^{x_{ij}x_{lj}x_{mj}} \left(1-\theta_4\right)^{(1-x_{ij})x_{lj}x_{mj}}. \qquad (2.30)
\end{aligned}
$$

Assim,

$$
\begin{aligned}
score\left(X_i, \mathbf{U}|D\right) &= \rho_{X_i}\left(\mathbf{U}\right) \int_0^1 \cdots \int_0^1 \left\{ \left(\prod_{v=1}^{4} \frac{1}{B\left(\alpha_v, \beta_v\right)} \theta_v^{\alpha_v-1} \left(1-\theta_v\right)^{\beta_v-1} \right) \times \right. \\
&\left. \times \left(\prod_{j=1}^{n} \Pr(X_{ij} = x_{ij}|X_{lj} = x_{lj}, X_{mj} = x_{mj}, \boldsymbol{\theta}_{X_i|\mathbf{U}}) \right) \right\} d\theta_1 \ldots d\theta_4
\end{aligned}
$$

$$
\begin{aligned}
score\left(X_i, \mathbf{U}|D\right) &= \frac{\rho_{X_i}\left(\mathbf{U}\right)}{\prod_{v=1}^{4} B\left(\alpha_v, \beta_v\right)} B\left(n_{(1,0,0)} + \alpha_1, n_{(0,0,0)} + \beta_1\right) \times \\
&\quad \times B\left(n_{(1,1,0)} + \alpha_2, n_{(0,1,0)} + \beta_2\right) B\left(n_{(1,0,1)} + \alpha_3, n_{(0,0,1)} + \beta_3\right) \times \\
&\quad \times B\left(n_{(1,1,1)} + \alpha_4, n_{(0,1,1)} + \beta_4\right), \qquad (2.31)
\end{aligned}
$$

onde $n_{(a,b,c)}$ = número de casos em D tal que $X_i = a$, $X_l = b$ e $X_m = c$, para $a = 0, 1$, $b = 0, 1$, $c = 0, 1$.

A determinação da função $score\left(X_i, \mathbf{U}|D\right)$ para conjuntos $\mathbf{U}$ com $3, 4, ...$ou k variáveis candidatas a serem pais de X_i é realizada de maneira semelhante à função $score\left(X_i, \mathbf{U}|D\right)$ para conjuntos $\mathbf{U}$ com $0, 1$ ou 2 variáveis descrita acima.

Definida $\pi(\prec |D)$, podemos construir a cadeia de Markov $\mathcal{M}$, com espaço de estados composto pelas $p!$ possíveis ordens, usando o algoritmo Metropolis-Hastings, no qual uma transição da ordem $\prec$ para a ordem $\prec'$ é aceita com probabilidade

$$
\alpha(\prec, \prec') = \min\left[1, \frac{\pi\left(\prec' |D\right)}{\pi\left(\prec |D\right)}\right],
$$

onde $\prec'$ é gerada a partir de $\prec$, através da inversão de duas variáveis escolhidas aleatoriamente e mantendo todas as outras na mesma posição

$$
\left(i_1 ... i_l ... i_m ... i_p\right) \longmapsto \left(i_1 ... i_m ... i_l ... i_p\right), \prec \rightarrow \prec'.
$$

Construída e simulada a cadeia de Markov, obtemos uma seqüência de ordens $\prec_1, ..., \prec_T$ e, assim, podemos aproximar o valor esperado de qualquer função $g(\prec)$ como

$$
E\left[g|D\right] \approx \frac{1}{T} \sum_{t=1}^{T} g\left(\prec_t\right).
$$

Especificamente, podemos assumir $g\left(\prec\right) = \Pr\left(f|\ \prec, D\right)$ para alguma característica f e computar $g\left(\prec_t\right) = \Pr\left(f|\ \prec_t, D\right)$.

A probabilidade *a posteriori* de um arco $X_i \longrightarrow X_l$, nosso maior interesse neste estudo, é aproximada por (2.20).

2.3.1 Aplicação em dados simulados

Para aplicação da metodologia de estimação da estrutura de dependência entre as variáveis de uma rede probabilística, consideramos um conjunto de variáveis $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$, cuja estrutura de dependência entre as variáveis é descrita na Figura 2.2 (a) e onde as variáveis X_i's são variáveis binárias com as seguintes distribuições,

$$X_1 \sim Bernoulli(\theta_{11}),$$

$$X_2 \sim Bernoulli(\theta_{21}),$$

$$X_3 \sim \begin{cases} Bernoulli(\theta_{31}), \text{ se } X_1 = X_2 = 0 \\ Bernoulli(\theta_{32}), \text{ se } X_1 = 0 \text{ e } X_2 = 1 \\ Bernoulli(\theta_{33}), \text{ se } X_1 = 1 \text{ e } X_2 = 0 \\ Bernoulli(\theta_{34}), \text{ se } X_1 = X_2 = 1 \end{cases},$$

$$X_4 \sim \begin{cases} Bernoulli(\theta_{41}), \text{ se } X_3 = 0 \\ Bernoulli(\theta_{42}), \text{ se } X_3 = 1 \end{cases} \text{ e}$$

$$X_5 \sim \begin{cases} Bernoulli(\theta_{51}), \text{ se } X_3 = 0 \\ Bernoulli(\theta_{52}), \text{ se } X_3 = 1 \end{cases},$$

onde $\theta_{il} = P(X_i = 1|\mathbf{pa}_G^l(X_i), G)$, $\mathbf{pa}_G^l(X_i)$ denota a l-ésima combinação dos valores assumidos por $\mathbf{Pa}_G(X_i)$, para $i = 1, ..., p$ e $l = 1, ..., q_i$ e onde q_i é o número de combinações dos valores de $\mathbf{Pa}_G(X_i)$.

Considerando $\pi(\prec)$ e $\pi(G|\prec)$ distribuições uniformes discretas, temos que

$$\pi(\prec|D) \propto \prod_{i=1}^{p} \sum_{\mathbf{U} \in \mathcal{U}_{i,\prec}} score\,(X_i, \mathbf{U}|D)\,, \tag{2.32}$$

onde $score\,(X_i, \mathbf{U}|D) = \int \left(\prod_{j=1}^{n} \Pr(X_{ij} = x_{ij}|\mathbf{u}_j, \boldsymbol{\theta}_{X_i|\mathbf{U}}) \right) \pi\left(\boldsymbol{\theta}_{X_i|\mathbf{U}}|G\right) d\boldsymbol{\theta}$.

Geramos a partir de $\boldsymbol{\theta} = (\theta_{11}, \theta_{21}, \theta_{31}, \theta_{32}, \theta_{33}, \theta_{34}, \theta_{41}, \theta_{42}, \theta_{51}, \theta_{52}) = (0.5, 0.3, 0.15, 0.25, 0.7, 0.85, 0.15, 0.9, 0.15, 0.7)$ 850 realizações da rede probabilística ilustrada na Figura 2.2 (a) e através do algoritmo Metropolis-Hastings, discutido na seção anterior, geramos

101000 ordens de $\pi(\prec|D)$. O número máximo considerado de pais por família, k, foi igual a 3; a ordem para a inicialização do algoritmo, $\prec^{(0)}$, foi uma ordenação aleatória das variáveis e assumimos distribuições *a priori* não informativas para todos os elementos de $\boldsymbol{\theta}_{X_i|\mathbf{U}}$, supondo $\alpha_v = \beta_v = 1$, para $1 \leq i \leq 5$ e $1 \leq v \leq 2^{|\mathbf{U}|}$.

Das 101000 ordens geradas pelo algoritmo, coletamos uma amostra de 2000 ordens. Para isto, descartamos as 1000 primeiras ordens geradas e coletamos uma a cada 50 realizações da cadeia de Markov gerada. Como $\mathbf{X}$ é composto por cinco variáveis ($p = 5$), temos $2\binom{p}{2} = 20$ possíveis arcos na estrutura e, para cada ordem da amostra, calculamos a probabilidade *a posteriori* $\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec_t)$ da presença de cada arco, definida em (2.20), para $i = 1, ..., 5$, $l = 1, ...5$ e $i \neq l$.

(a) (b)

Figura 2.2: (a) Estrutura de $\mathbf{X}$. (b) Estrutura estimada de $\mathbf{X}$.

Finalmente, a probabilidade do arco $X_i \rightarrow X_l$ estar presente na estrutura de $\mathbf{X}$ pode ser estimada por

$$\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec) = \frac{\sum_{t=1}^{2000} \widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec_t)}{2000}. \tag{2.33}$$

Para o conjunto de dados considerado neste estudo, as probabilidades estimadas de presença dos arcos na estrutura são apresentadas na matriz P a seguir,

$$P = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & X_2 & X_3 & X_4 & X_5 \end{array} \\ \begin{pmatrix} & 0.150 & 0.237 & 0.011 & 0.018 \\ 0.078 & & 0.232 & 0.004 & 0.007 \\ 0.763 & 0.732 & & 0.866 & 0.891 \\ 0.022 & 0.032 & 0.134 & & 0.200 \\ 0.012 & 0.013 & 0.109 & 0.078 & \end{pmatrix} \end{array},$$

onde $p_{il} = \widehat{\Pr}\left(X_i \in \mathbf{Pa}_G(X_l)|D, \prec\right)$.

Fixando uma probabilidade de corte igual a 0.50, podemos construir a estrutura de dependência de $\mathbf{X} = (X_1, X_2, ..., X_5)$ como sendo composta pelos arcos com probabilidade *a posteriori* estimada superior a 0.50. Desta maneira, a estrutura estimada para $\mathbf{X}$ possui os arcos $X_3 \rightarrow X_1$, $X_3 \rightarrow X_2$, $X_3 \rightarrow X_4$ e $X_3 \rightarrow X_5$, ou seja, as variáveis X_1, X_2, X_4 e X_5 são diretamente dependentes de X_3.

Comparando a estrutura estimada, representada na Figura 2.2 (b), com a estrutura utilizada para gerar os dados, representada na Figura 2.2 (a), observamos que os arcos $X_3 \rightarrow X_4$ e $X_3 \rightarrow X_5$ foram corretamente estimados, ao passo que os arcos $X_1 \rightarrow X_3$ e $X_2 \rightarrow X_3$ foram invertidos.

Esta inversão na estimação dos arcos $X_1 \rightarrow X_3$ e $X_2 \rightarrow X_3$, que são justamente os arcos que associam os dois vértices raiz X_1 e X_2 (variáveis independentes) ao vértice filho, X_3, comum a ambos, nos indica a existência de uma possível deficiência da metodologia em identificar as variáveis independentes (vértices raiz que iniciam a rede e que influenciam, diretamente ou indiretamente, na distribuição de probabilidades das demais variáveis da rede) e estimar seus vértices filhos. Já a estrutura de dependência entre variáveis que não são nós raiz (no exemplo, X_3, X_4 e X_5) é corretamente estimada. Assim, a metodologia apresentada neste estudo parece estimar como variáveis independentes os filhos dos verdadeiros vértices raiz e identificar as verdadeiras variáveis independentes como dependentes dos nós raiz estimados.

A convergência da cadeia de Markov para a distribuição *a posteriori* de interesse pode ser verificada graficamente, através do gráfico do valor gerado a cada iteração do algoritmo *versus* a ordem de geração dos valores. Na situação considerada neste estudo, a cada iteração uma seqüência de variáveis (uma ordem) é gerada. Portanto, não é possível

construir um gráfico plano dos valores gerados *versus* ordem de geração.

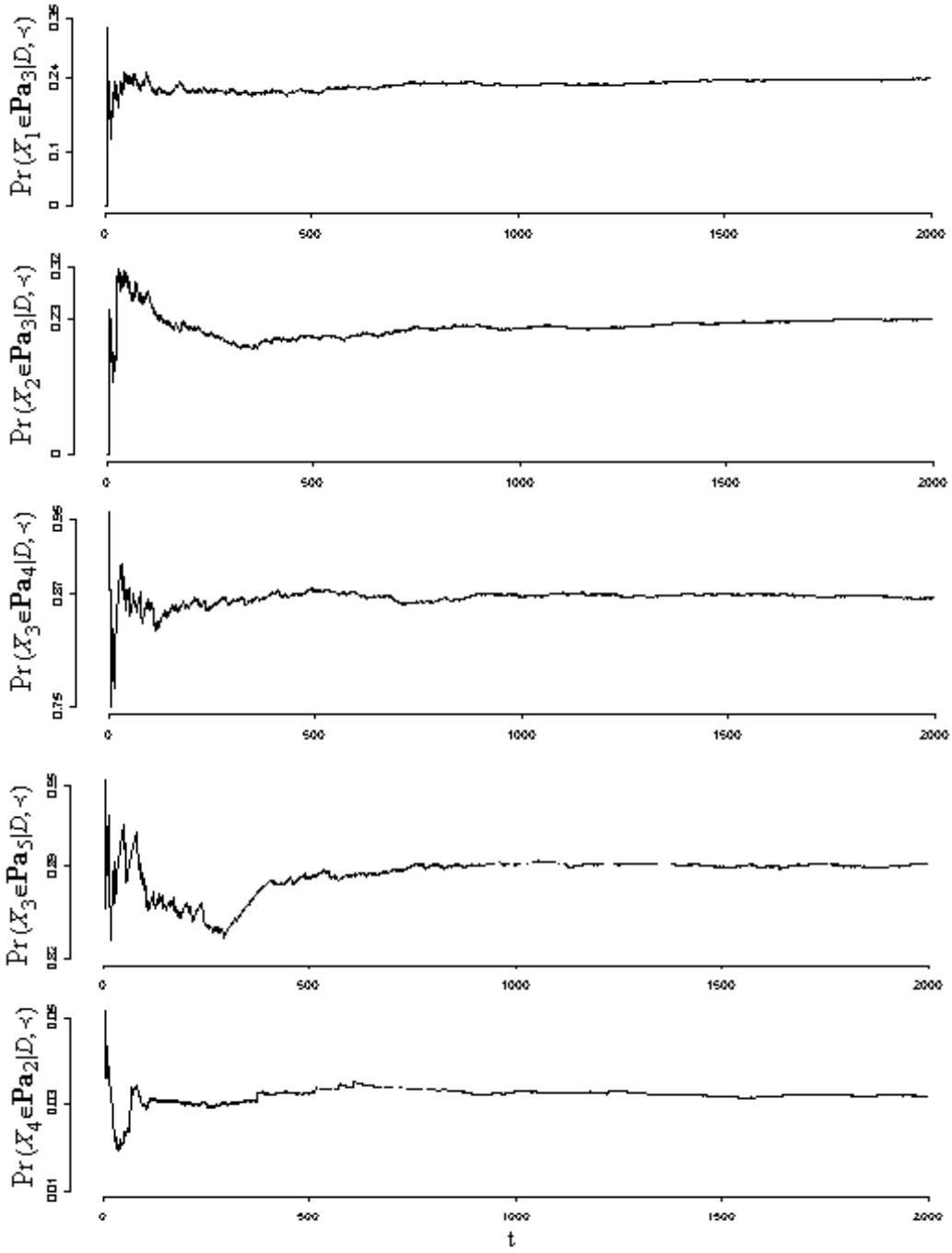


Figura 2.3: Gráficos de Convergência, onde $\mathbf{Pa}_l = \mathbf{Pa}_G(X_l)$

Por este motivo, optamos, para verificar a convergência da cadeia para a distribuição *a posteriori* da ordem, $\pi(\prec|D)$, por analisar o comportamento de $\widehat{\Pr}\left(X_i \in \mathbf{Pa}_G(X_l)|D, \prec\right)$, para $i = 1, .., 5$, $l = 1, ..., 5$ e $i \neq l$, a cada nova ordem considerada no estudo.

Como existem 20 possíveis arcos na estrutura de $\mathbf{X} = (X_1, ..., X_5)$, consideramos

na análise da convergência apenas o comportamento de $\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec)$ para cinco arcos, dois deles altamente prováveis e corretamente estimados na estrutura ($X_3 \rightarrow X_4$ e $X_3 \rightarrow X_5$), dois que foram estimados erroneamente ($X_1 \rightarrow X_3$ e $X_2 \rightarrow X_3$) e um arco pouco provável ($X_4 \rightarrow X_2$) . Para cada um dos cinco arcos considerados na análise, construímos gráficos da $\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec)$ parcial calculada em cada ordem da amostra. A probabilidade parcial da t-ésima ordem é calculada como em (2.33), levando em consideração apenas a t-ésima amostra e as ordens anteriores a ela e não todas as 2000 ordens.

Assim, verificamos a convergência da cadeia para a distribuição *a posteriori* da ordem através da convergência das probabilidades parciais de cada arco para a sua respectiva $\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec)$. Na Figura 2.3 , observamos que, com a evolução da cadeia, as probabilidades parciais de cada arco convergem para $\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec)$, pois, a partir da ordem 500, praticamente todas as probabilidades parciais em cada ordem se estabilizam em $\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec)$. Desta maneira, concluímos que a cadeia de Markov $\mathcal{M}$ gerada pelo algoritmo convergiu para a distribuição *a posteriori* $\pi(\prec |D)$.

2.3.2 Avaliação dos estimadores

Na estimação da estrutura de dependência do conjunto de variáveis considerada na seção anterior e ilustrada na Figura 2.2 (a), observamos a possível existência de uma deficiência da metodologia apresentada neste estudo em identificar as variáveis independentes e estimar as suas variáveis filhas. Desta maneira, com o objetivo de verificar se esta má estimação das variáveis independentes é um problema da metodologia em estudo ou se foi ocasionada apenas devido ao conjunto de variáveis considerada na seção anterior, realizamos um estudo de simulação baseado no conjunto de variáveis binárias $\mathbf{X} = (X_1, X_2, ..., X_7)$ com quatro diferentes estruturas de dependência que possuem distintas variáveis como vértices raiz.

Geramos a partir de $\boldsymbol{\theta}_1$ =(0.3, 0.5, 0.3, 0.6, 0.6, 0.3, 0.75, 0.25, 0.3, 0.85, 0.15, 0.9, 0.15, 0.70), $\boldsymbol{\theta}_2$ =(0.3, 0.5, 0.6, 0.3, 0.75, 0.25, 0.3, 0.85, 0.15, 0.9, 0.15, 0.70), $\boldsymbol{\theta}_3$ =(0.3, 0.5, 0.3, 0.6, 0.6, 0.3, 0.3, 0.85, 0.15, 0.9, 0.15, 0.70) e $\boldsymbol{\theta}_4$ =(0.3, 0.5, 0.3, 0.6, 0.6, 0.3, 0.85, 0.15, 0.9, 0.15, 0.70) 850 realizações das redes probabilísticas ilustradas, respectivamente, nas Figuras 2.4 (a), 2.5 (a), 2.6 (a) e 2.7(a) e, para cada rede probabilística,

geramos através do algoritmo Metropolis-Hastings 251000 ordens de $\pi(\prec|D)$. O número máximo considerado de pais por família, k, foi igual a 3; a ordem para a inicialização do algoritmo,$\prec^{(0)}$, foi uma ordenação aleatória das variáveis e assumimos distribuições *a priori* não informativas para todos os elementos de $\boldsymbol{\theta}_{X_i|\mathbf{U}}$, supondo $\alpha_v = \beta_v = 1$, para $1 \leq i \leq 7$ e $1 \leq v \leq 2^{|\mathbf{U}|}$.

Das 251000 ordens geradas pelo algoritmo para cada rede, coletamos uma amostra de 5000 ordens. Para isto, descartando as 1000 primeiras ordens geradas e coletamos uma a cada 50 ordens. Como $\mathbf{X}$ é composto por sete variáveis ($p = 7$), temos $2\binom{p}{2} = 42$ possíveis arcos na estrutura e, para cada ordem da amostra, estimamos a probabilidade *a posteriori* $\Pr(X_i \in \mathbf{Pa}_G(X_l)|D, \prec_t)$ da presença de cada arco, definida em (2.20), para $i = 1, ..., 7$, $l = 1, ...7$ e $i \neq l$.

Finalmente, a probabilidade do arco $X_i \rightarrow X_l$ estar presente na estrutura de $\mathbf{X}$ pode ser calculada como

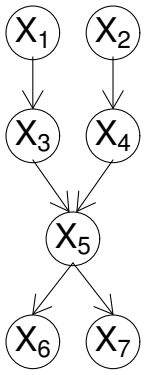
$$\widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec) = \frac{\sum_{t=1}^{5000} \widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec_t)}{5000}. \tag{2.34}$$

Fixando uma probabilidade de corte igual a 0.50, podemos construir a estrutura de dependência de $\mathbf{X}$ como sendo composta pelos arcos com probabilidade estimada superior a 0.50. As probabilidades estimadas de presença dos arcos nas estruturas são apresentadas nas matrizes P_1, P_2, P_3 e P_4 nas Figuras 2.4 (c), 2.5 (c), 2.6 (c) e 2.7(c), respectivamante, onde $p_{il} = \widehat{\Pr}(X_i \in \mathbf{Pa}_G(X_l)|D, \prec)$. As estruturas estimadas são representadas nas Figuras 2.4 (b), 2.5 (b), 2.6 (b) e 2.7(b).

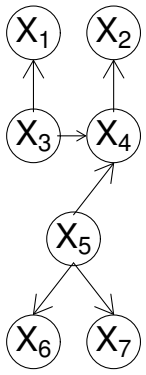
Comparando as estruturas estimadas com as estruturas utilizadas para a geração dos dados, observamos que, geralmente, a relação de dependência entre os vértices raiz e seus filhos foi invertida na estimação, ou seja, os verdadeiros vértices raiz foram estimados como dependentes das variáveis que, na realidade, são suas filhas.

Em algumas estruturas, a inversão entre vértices pais e filhos se repete entre as variáveis filhas dos vértices raiz (variáveis da segunda geração na rede) e seus filhos (variáveis da terceira geração). No entanto, observamos nas matrizes de probabilidades estimadas dos arcos que a probabilidade dos arcos que invertem a verdadeira relação de dependência entre variáveis da segunda e terceira geração, apesar de serem superiores

a 0.50, não são tão altas quanto as probabilidades dos arcos que invertem a verdadeira relação de associação entre os vértices raiz e seus filhos e variam, geralmente, entre 0.50 e 0.60. Já as probabilidades estimadas dos arcos que representam a verdadeira relação de dependência entre as variáveis da segunda e terceira geração são próximas a 0.50, apesar de serem inferiores, e variam, usualmente, entre 0.40 e 0.50.



(*a*) Estrutura de $\mathbf{X}$



(*b*) Estrutura estimada

$$P_1 = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{array} \\ \begin{pmatrix} & 0.015 & 0.274 & 0.008 & 0.012 & 0.009 & 0.013 \\ 0.006 & & 0.007 & 0.096 & 0.001 & 0.009 & 0.006 \\ 0.726 & 0.021 & & 0.589 & 0.393 & 0.009 & 0.044 \\ 0.025 & 0.904 & 0.076 & & 0.388 & 0.005 & 0.010 \\ 0.017 & 0.014 & 0.089 & 0.603 & & 0.883 & 0.857 \\ 0.013 & 0.009 & 0.008 & 0.003 & 0.117 & & 0.072 \\ 0.006 & 0.011 & 0.043 & 0.004 & 0.143 & 0.036 & \end{pmatrix} \end{array}$$

(*c*) Matriz de Probabilidades Estimadas dos Arcos

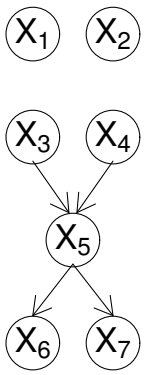
Figura 2.4. Rede Probabilística 1

Já a estrutura de dependência entre variáveis alocadas a partir da terceira geração no grau de parentesco é, geralmente, estimada corretamente.

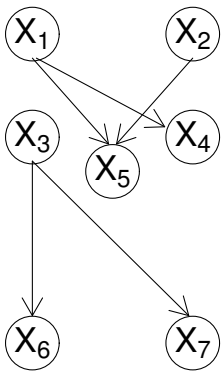
Na estrutura representada na Figura 2.4, observamos que os arcos $X_1 \rightarrow X_3$ e $X_2 \rightarrow X_4$, associados às variáveis independentes X_1 e X_2, foram invertidos, assim como o

arco que associa X_4 a X_5. O arco $X_3 \rightarrow X_4$ foi acrescentado à estrutura e $X_3 \rightarrow X_5$ não foi evidente. Já os arcos $X_5 \rightarrow X_6$ e $X_5 \rightarrow X_7$ foram corretamente estimados.

A Figura 2.5 apresenta uma estrutura com duas variáveis, no início da rede, totalmente independentes das demais variáveis, ou seja, sem alguma associação com os outros vértices do conjunto. A metodologia identificou dois subconjuntos independentes de variáveis dentro do conjunto $\mathbf{X}$. O primeiro subconjunto é composto por X_1, X_2, X_4 e X_5 e o segundo por X_3, X_6 e X_7. A dependência de X_6 e X_7 com X_3, mesmo que seja indireta e ocorra através de X_5, existe na estrutura verdadeira, mas a associação estimada entre X_1, X_2, X_4 e X_5 é incorreta. As probabilidades das distribuições das variáveis utilizadas para a geração dos dados, $\boldsymbol{\theta}_2$, podem ser o motivo desta errada associação.



(a) Estrutura de $\mathbf{X}$



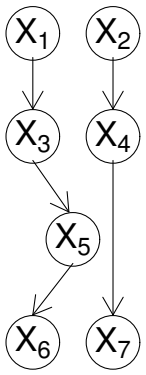
(b) Estrutura estimada

$$P_2 = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{array} \\ \begin{pmatrix} & 0.352 & 0.021 & 0.640 & 0.709 & 0.040 & 0.204 \\ 0.023 & & 0.012 & 0.024 & 0.677 & 0.011 & 0.013 \\ 0.024 & 0.032 & & 0.020 & 0.010 & 0.777 & 0.640 \\ 0.281 & 0.030 & 0.014 & & 0.022 & 0.021 & 0.016 \\ 0.026 & 0.323 & 0.007 & 0.037 & & 0.010 & 0.005 \\ 0.042 & 0.025 & 0.223 & 0.013 & 0.014 & & 0.012 \\ 0.215 & 0.034 & 0.360 & 0.028 & 0.009 & 0.008 & \end{pmatrix} \end{array}$$

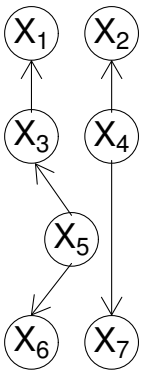
(c) Matriz de Probabilidades Estimadas dos Arcos

Figura 2.5. Rede Probabilística 2

A rede probabilística representada na Figura 2.6 possui dois subconjuntos independentes de variáveis. O primeiro subconjunto é formado pelas variáveis X_1, X_3, X_5 e X_6 e o segundo por X_2, X_4 e X_7. Estes subconjuntos independentes foram corretamente estimados, mas os arcos que associam os vértices raiz aos seus filhos foram invertidos. O arco $X_3 \rightarrow X_5$ também foi invertido, no entanto, a sua pro- babilidade estimada e a probabilidade estimada do arco inverso $X_5 \rightarrow X_3$ são muito próximas. Este fato evidencia que a metodologia identificou a associação entre X_3 e X_5, mas não percebeu precisamente qual é a variável dependente.



(a) Estrutura de $\mathbf{X}$



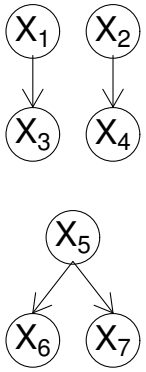
(b) Estrutura estimada

$$P_3 = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \end{array} \\ \begin{pmatrix} & 0.012 & 0.159 & 0.013 & 0.008 & 0.004 & 0.009 \\ 0.012 & & 0.009 & 0.232 & 0.008 & 0.005 & 0.018 \\ 0.841 & 0.031 & & 0.024 & 0.496 & 0.012 & 0.015 \\ 0.013 & 0.768 & 0.017 & & 0.020 & 0.073 & 0.733 \\ 0.019 & 0.028 & 0.504 & 0.024 & & 0.864 & 0.024 \\ 0.007 & 0.117 & 0.008 & 0.048 & 0.136 & & 0.015 \\ 0.014 & 0.030 & 0.010 & 0.267 & 0.014 & 0.009 & \end{pmatrix} \end{array}$$

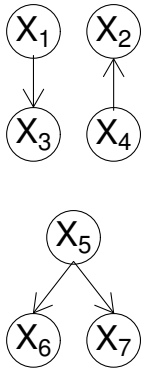
(c) Matriz de Probabilidades Estimadas dos Arcos

Figura 2.6. Rede Probabilística 3

A Figura 2.7 também apresenta uma rede probabilística com subconjuntos independentes de variáveis. Nesta estrutura, o primeiro subconjunto é composto por X_1 e X_3, o segundo por X_2 e X_4 e o terceiro pelas variáveis X_5, X_6 e X_7. Estes três subconjuntos foram corretamente estimados, assim como a estrutura de dependência entre as variáveis do primeiro e terceiro subconjunto. Já a relação entre X_2 e X_4 foi invertida, pois a probabilidade estimada do arco $X_4 \rightarrow X_2$ é superior a 0.50. No entanto, a probabilidade estimada de $X_2 \rightarrow X_4$, apesar de ser inferior, é muito próxima a 0.50.



(a) Estrutura de $\mathbf{X}$



(b) Estrutura estimada

$P_4 =$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1		0.021	0.559	0.021	0.010	0.010	0.015
X_2	0.021		0.024	0.491	0.012	0.008	0.059
X_3	0.441	0.020		0.016	0.010	0.009	0.014
X_4	0.029	0.509	0.029		0.008	0.007	0.015
X_5	0.151	0.061	0.229	0.044		0.748	0.827
X_6	0.105	0.027	0.125	0.023	0.252		0.029
X_7	0.025	0.124	0.033	0.066	0.173	0.014	

(c) Matriz de Probabilidades Estimadas dos Arcos

Figura 2.7. Rede Probabilística 4

Capítulo 3

Modelos Markovianos Ocultos

As redes probabilísticas também descrevem processos estocásticos complexos, especificando de uma maneira simples as independências condicionais e fornecendo uma parametrização compacta do modelo. Os modelos Markovianos ocultos, HMMs (do inglês *Hidden Markov Models*), são alguns dos modelos usualmente apresentados na forma de grafos devido à complexidade da notação necessária para apresentá-los na forma analítica.

Neste capítulo, definimos alguns modelos Markovianos ocultos distintos e apresentamos os estimadores para os parâmetros envolvidos em cada modelo.

3.1 Modelos Markovianos ocultos

O HMM é um caso particular das redes probabilísticas dinâmicas e pode ser definido como um processo estocástico finito, no qual cada estado está associado a uma distribuição de probabilidades. As transições entre os estados deste processo são governadas por um conjunto de probabilidades denominado probabilidades de transição e, em cada estado, um resultado ou uma observação é gerada de acordo com a distribuição de probabilidades associada. Somente o resultado, não o estado, é visível a um "observador" e, portanto, os estados são "ocultos" ao exterior; daí o nome modelo Markoviano oculto. Desta maneira, este modelo é caracterizado pelos três conjuntos de probabilidades, ou seja, pelos conjuntos das probabilidades iniciais dos estados ocultos, probabilidades de transição entre os estados ocultos e probabilidades condicionais.

Na modelagem de um HMM estamos interessados no conjunto de variáveis aleatórias

$$\mathbf{Z} = \{S_1, Y_1, S_2, Y_2, ..., S_{T-1}, Y_{T-1}, S_T, Y_T\},$$

onde S_t é a variável oculta no tempo t e Y_t é a variável observada, que pode ser discreta ou contínua, no tempo t, para $1 \leq t \leq T$.

Seja Y_t uma variável discreta, podemos definir os seguintes elementos comuns a diferentes HMMs:

N : número de estados ocultos no modelo;

M : número de resultados observáveis (observações) em cada estado oculto;

T : tamanho da seqüência de observações, ou seja, número de resultados observados;

$\mathbf{S} = S_1, S_2, ..., S_T$: seqüência de variáveis ocultas, onde $S_t \in \{1, 2, ..., N\}$ denota o estado no tempo t;

$\mathbf{Y} = Y_1, Y_2, ..., Y_T$: seqüência de variáveis observáveis, onde $Y_t \in \{1, 2, ..., M\}$ denota o resultado observável no tempo t;

$\mathbf{s} = s_1, s_2, ..., s_T$: realização do processo aleatório $\mathbf{S} = S_1, S_2, ..., S_T$ e;

$\mathbf{y} = y_1, y_2, ..., y_T$: realização do processo aleatório $\mathbf{Y} = Y_1, Y_2, ..., Y_T$.

A utilização de HMMs consiste, basicamente, na resolução de três problemas principais. Um dos problemas de interesse é o cálculo da probabilidade de ocorrência da seqüência de observações dado que a estrutura e os parâmetros das distribuições envolvidas no modelo são conhecidos, ou seja, $\Pr(\mathbf{Y} = \mathbf{y}|$ modelo). Outro problema envolve a identificação da seqüência de estados ocultos mais provável dada a seqüência de observações e o modelo. Por sua vez, a estimação dos parâmetros envolvidos no modelo (probabilidades inciais dos estados, probabilidades de transição entre os estados e os parâmetros envolvidos nas distribuições de probabilidades condicionais) é o outro interesse no contexto do HMM.

Estes problemas, principalmente a estimação dos parâmetros, podem ser resolvidos pelo critério da máxima verossimilhança ou através de métodos bayesianos.

Por simplicidade de notação, o HMM com dependência de ordem r em $\mathbf{S}$ e de ordem q em $\mathbf{Y}$ será representado por HMM(r, q), para $r = 0, 1, 2, ...$ e $q = 0, 1, 2, ...$

3.2 HMM com dependência de Markov de primeira ordem entre os estados ocultos, HMM$(1,0)$

O HMM mais simples é o modelo que assume dependência de Markov de primeira ordem somente entre os estados ocultos e que obedece as seguintes relações de independência condicional,

$$S_t \perp \{S_1, Y_1, ..., S_{t-2}, Y_{t-2}, Y_{t-1}\} | S_{t-1},\ 2 \leq t \leq T \text{ e} \tag{3.1}$$

$$Y_t \perp \{S_1, Y_1, ..., S_{t-1}, Y_{t-1}\} | S_t,\ 2 \leq t \leq T\ . \tag{3.2}$$

A estrutura de dependência das variáveis deste HMM, assim como o grafo moral desta estrutura, são representados, respectivamente, nas Figuras 3.1 (a) e 3.1 (b). Naturalmente, a direção dos arcos entre S_{t-1} e S_t vai de S_{t-1} a S_t e a direção dos arcos entre S_t e Y_t vai de S_t a Y_t. A posição dos arcos implica que Y_t é condicionalmente independente de S_{t-1} dado S_t e S_t é condicionalmente independente de S_{t-2} dado S_{t-1}.

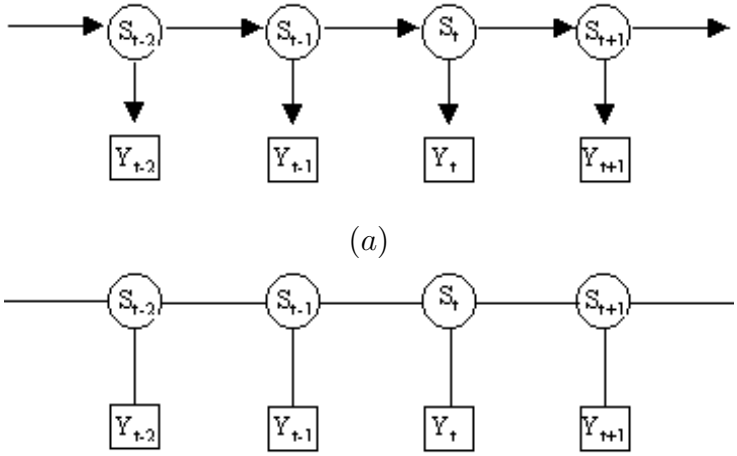


Figura 3.1: (a) Estrutura do HMM$(1,0)$. (b) Grafo moral da estrutura

Um HMM com dependência de Markov de primeira ordem somente entre os estados ocultos é especificado pelos seguintes elementos:

$\pi = \{\pi_k\}$, onde $\pi_k = \Pr(S_1 = k|\pi)$ é a probabilidade de estar no estado k no início do experimento, ou seja, em $t = 1$, para $k = 1, 2, ..., N$;

$A = \{a_{kl}\}$, representa a matriz com as probabilidades de transição, onde

$$a_{kl} = \Pr(S_{t+1} = l | S_t = k, A) \tag{3.3}$$

é a probabilidade de estar no estado l no tempo $t+1$ dado que no instante t o estado era k, para $1 \leq k, l \leq N$;

$P = \{p_{kj}\}$, onde

$$p_{kj} = \Pr(Y_t = j | S_t = k, P) \tag{3.4}$$

é a probabilidade de observar o resultado j no instante t dado que a cadeia oculta está no estado k, para $k = 1, 2, ..., N$ e $j = 1, 2, ..., M$ e;

$\lambda = (A, P, \pi)$ é a notação compacta para denotar o HMM.

Uma maneira direta de calcular $\Pr(\mathbf{Y} = \mathbf{y}|\lambda)$ é determinar $\Pr(\mathbf{Y} = \mathbf{y}|\ \mathbf{s}, \lambda)$ para uma seqüência de estados fixa $\mathbf{s}$, multiplicá-la pela probabilidade de $\mathbf{s}$ dado o modelo, $\Pr(\mathbf{S} = \mathbf{s}|\lambda)$, e então somá-la para todas as possíveis seqüências de estados $\mathbf{s}$ de tamanho T. Portanto,

$$\begin{aligned}
\Pr(\mathbf{Y} &= \mathbf{y}|\lambda) = \sum_{s_1,\ldots,s_T} \Pr(\mathbf{Y} = \mathbf{y}|\mathbf{s}, \lambda)\ \Pr(\mathbf{S} = \mathbf{s}|\lambda) \\
&= \sum_{s_1,\ldots,s_T} \left(\begin{array}{c} \Pr(S_1 = s_1|\lambda) \Pr(Y_1 = y_1|s_1, \lambda) \times \\ \times \left(\prod_{t=2}^{T} \Pr(S_t = s_t|s_{t-1}, \lambda) \Pr(Y_t = y_t|s_t, \lambda) \right) \end{array} \right) \\
&= \sum_{s_1,\ldots,s_T} \pi_{s_1} p_{s_1 y_1} \left(\prod_{t=2}^{T} a_{s_{t-1} s_t} p_{s_t y_t} \right),
\end{aligned} \tag{3.5}$$

onde s_t denota os possíveis valores da variável oculta S_t e y_t denota o valor observado da variável Y_t, para $t = 1, 2, ...T$.

Este cálculo exige computações da ordem de $2TN^T$, que dificulta sua utilização em muitas aplicações. Na prática, o passo *forward* do algoritmo *forward-backward* (Poritz, 1988; Rabiner, 1989; Dugad e Desai, 1996), que calcula a probabilidade da seqüência observada, realiza este cálculo inferencial com menos complexidade e computações na ordem de TN^2.

O problema de identificação da seqüência de estados ocultos mais provável pode ser solucionado através do algoritmo de Viterbi (Poritz, 1988; Rabiner, 1989; Dugad e Desai, 1996). Este é um método eficiente e recursivo, com complexidade na ordem de TN^2, assim como o passo *forward* do algoritmo *forward-backward*.

Já as estimativas dos parâmetros A, P e π deste HMM podem ser obtidas através da maximização da função de verossimilhança definida em (3.5) em relação a A, P e π. A maximização da função (3.5) pode ser realizada através do algoritmo EM, também conhecido como "Fórmulas de re-estimação de Baum-Welch"(Dugad e Desai, 1996) para este problema particular.

Zuanetti e Milan (2003) estudaram a performance destes estimadores de máxima verossimilhança em diversas situações.

3.3 HMM com dependência de Markov de primeira ordem entre os estados ocultos e os observáveis, HMM$(1, 1)$

Um HMM mais sofisticado do que o apresentado na seção anterior e descrito em Boys *et al.* (2000), é o modelo cuja seqüência de variáveis observáveis também é considerada uma cadeia de Markov de primeira ordem, ou seja, assumimos que transições entre os estados observáveis (resultados) $Y_{t-1} \rightarrow Y_t$ seguem um processo estocástico de primeira ordem, cuja escolha da matriz de transição é determinada pelo estado oculto S_t.

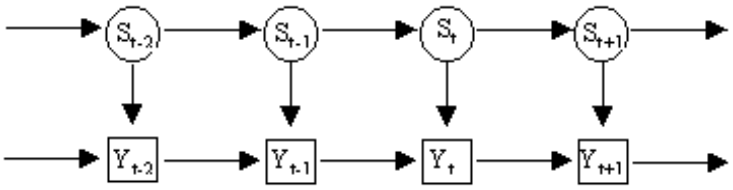


Figura 3.2: Grafo Direcional Acíclico do HMM$(1, 1)$

Uma maneira de entender a estrutura de dependência do modelo é através do grafo direcional acíclico, representado na Figura 3.2, no qual podemos observar que o modelo obedece as seguintes relações de independência condicional,

$$S_t \perp \{S_1, Y_1, ..., S_{t-2}, Y_{t-2}, Y_{t-1}\} | S_{t-1},\ 2 \leq t \leq T \text{ e} \tag{3.6}$$

$$Y_t \perp \{S_1, Y_1, ..., S_{t-2}, Y_{t-2}, S_{t-1}\} | S_t \text{ e } Y_{t-1},\ 2 \leq t \leq T\ . \tag{3.7}$$

Assumindo que Y_1 e S_1 seguem distribuições uniformes discretas independentes, este HMM é especificado pelos seguintes elementos:

$A = \{a_{kl}\}$, representa a matriz com as probabilidades de transição entre os estados ocultos, onde

$$a_{kl} = \Pr(S_{t+1} = l | S_t = k, A) \tag{3.8}$$

é a probabilidade de estar no estado l no tempo $t+1$ dado que no instante t o estado era k, para $1 \leq k, l \leq N$ e;

$\mathbf{P} = \left\{P^{(1)}, P^{(2)}, ..., P^{(N)}\right\}$, onde $P^{(k)} = \left\{p_{ij}^{(k)}\right\}$, representa a matriz com as probabilidades de transição entre os estados observáveis (resultados) associada ao k-ésimo estado oculto, e

$$p_{ij}^{(k)} = \Pr(Y_t = j | S_t = k, Y_{t-1} = i, \mathbf{P}) \tag{3.9}$$

é a probabilidade de observar o resultado j no instante t dado que a cadeia oculta está no estado k e no instante anterior, $t-1$, foi observado o resultado i, para $k = 1, 2, ..., N$ e $1 \leq i, j \leq M$.

Assim, a função de verossimilhança para os parâmetros $\mathbf{P}$ e A do modelo, dada a seqüência observada $\mathbf{y}$ e a não observada $\mathbf{s}$, é

$$\begin{aligned}
L(\mathbf{P}, A | \mathbf{y}, \mathbf{s}) &= \Pr(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s} | A, \mathbf{P}) \\
&= \Pr(S_1 = s_1 | A, \mathbf{P}) \Pr(Y_1 = y_1 | s_1, A, \mathbf{P}) \Pr(S_2 = s_2 | s_1, y_1, A, \mathbf{P}) \times \\
\times \Pr(Y_2 &= y_2 | s_1, y_1, s_2, A, \mathbf{P}) ... \Pr(S_T = s_T | s_1, y_1, ..., s_{T-1}, y_{T-1}, A, \mathbf{P}) \times \\
\times \Pr(Y_T &= y_T | \mathbf{s}, y_1, ..., y_{T-1}, A, \mathbf{P}).
\end{aligned}$$

Pelas relações de independência condicional e suposições do modelo,

$$
\begin{aligned}
L(\mathbf{P}, A|\mathbf{y}, \mathbf{s}) &= \Pr(S_1 = s_1|A, \mathbf{P}) \Pr(Y_1 = y_1|A, \mathbf{P}) \times \\
&\quad \times \left(\prod_{t=2}^{T} \Pr(S_t = s_t|s_{t-1}, A, \mathbf{P}) \Pr(Y_t = y_t|y_{t-1}, s_t, A, \mathbf{P}) \right) \\
&= \frac{1}{NM} \prod_{t=2}^{T} a_{s_{t-1}s_t} p^{(s_t)}_{y_{t-1}y_t} \\
&= \frac{1}{NM} \left(\prod_{t=2}^{T} a_{s_{t-1}s_t} \right) \left(\prod_{t=2}^{T} p^{(s_t)}_{y_{t-1}y_t} \right) \\
&= \frac{1}{NM} \left(\prod_{k=1}^{N} \prod_{l=1}^{N} a_{kl}^{m_{kl}} \right) \left(\prod_{k=1}^{N} \prod_{i=1}^{M} \prod_{j=1}^{M} \left(p^{(k)}_{ij} \right)^{n^{(k)}_{ij}} \right), \qquad (3.10)
\end{aligned}
$$

onde

$$
m_{kl} = \sum_{t=1}^{T-1} I\left(s_t = k, s_{t+1} = l\right) \text{ e}
$$

$$
n^{(k)}_{ij} = \sum_{t=2}^{T} I\left(y_{t-1} = i, y_t = j, s_t = k\right),
$$

onde $I(A) = \begin{cases} 1 \text{ se } A \text{ é verdade} \\ 0 \text{ caso contrário} \end{cases}$.

Neste HMM, as inferências sobre A e $\mathbf{P}$ podem ser realizadas através de métodos bayesianos com simulação da distribuição *a posteriori* de A e $\mathbf{P}$ usando o método Monte Carlo em cadeias de Markov (MCMC), assim como a identificação dos estados ocultos, que são tratados como variáveis não observadas e são simuladas da distribuição condicional.

3.3.1 Distribuições *a priori*

Considere $\mathbf{p}_i = \{p_{ij}\}$ uma linha de uma das matrizes de transição dos estados observáveis, para $j = 1, 2, ..., M$. Dado a forma multinomial da função de verossimilhança, a distribuição *a priori* conjugada para o vetor M-dimensional $\mathbf{p}_i$ é uma distribuição Dirichlet definida no simplex, com densidade

$$\pi\left(\mathbf{p}_i\right) \propto \prod_{j=1}^{M} p_{ij}^{\alpha_{ij}-1},$$

onde $0 < p_{ij} < 1$, $j = 1, 2, ..., M$, $\sum_{j=1}^{M} p_{ij} = 1$ e $\boldsymbol{\alpha}_i = \{\alpha_{ij}\}$ são os parâmetros positivos da distribuição.

Supondo que as M linhas das N matrizes de transição entre os estados observáveis são independentes e que

$$\mathbf{p}_i^{(k)} = \{p_{ij}^{(k)}\} \sim \mathcal{D}(\boldsymbol{\alpha}_i^{(k)}),$$

para $k = 1, 2, ..., N$ e $i = 1, 2, ..., M$, temos que a distribuição *a priori* para $\mathbf{P}$ tem densidade

$$\begin{aligned} \pi\left(\mathbf{P}\right) &\propto \prod_{k=1}^{N}\prod_{i=1}^{M} \pi\left(\mathbf{p}_i^{(k)}\right) \\ &= \prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M} \left(p_{ij}^{(k)}\right)^{\alpha_{ij}^{(k)}-1}. \end{aligned} \tag{3.11}$$

Similarmente, usamos a distribuição Dirichlet N-dimensional como a distribuição *a priori* para as linhas $\mathbf{a}_k = \{a_{kl}\}$ da matriz de transição entre os estados ocultos A. Assumindo que N linhas da matriz A são independentes e que

$$\mathbf{a}_k = \{a_{kl}\} \sim \mathcal{D}(\boldsymbol{\beta}_k),$$

para $1 \leq k, l \leq N$ e onde $\boldsymbol{\beta}_k = \{\beta_{kl}\}$ são os parâmetros positivos da distribuição, temos que a distribuição *a priori* para A tem densidade

$$\begin{aligned} \pi\left(A\right) &\propto \prod_{k=1}^{N} \pi\left(\mathbf{a}_k\right) \\ &= \prod_{k=1}^{N}\prod_{l=1}^{N} a_{kl}^{\beta_{kl}-1}. \end{aligned} \tag{3.12}$$

Vale destacar que as componentes univariadas da distribuição Dirichlet seguem

uma distribuição Beta e são negativamente correlacionadas, ou seja, se $(X_1, X_2, ..., X_p) \sim \mathcal{D}(\delta_1, \delta_2, ..., \delta_p)$ com $\Delta = \sum_{i=1}^{p} \delta_i$, então

$$\begin{aligned} X_i &\sim \text{Beta } (\delta_i, \Delta - \delta_i) \text{ e} \\ \text{Cor}(X_i, X_j) &= -\sqrt{\frac{\delta_i \delta_j}{(\Delta - \delta_i)(\Delta - \delta_j)}}, \end{aligned}$$

para $1 \leq i, j \leq p$ e $i \neq j$. Este resultado pode ser útil na obtenção dos parâmetros da distribuição *a priori*.

Outras possíveis distribuições *a priori* que permitem maior flexibilidade na estrutura de covariância são a distribuição Normal Logística e a distribuição de Aitchison (Aitchison, 1986), que inclui as distribuições Dirichlet e Normal Logística como casos especiais. Uma desvantagem do uso da distribuição Normal Logística é o fato dela não ser conjugada para o modelo Multinomial, o que dificulta, mas não impossibilita, o uso do método MCMC. A distribuição de Aitchison apresenta problemas na especificação de seus parâmetros, que apresentam estrutura analiticamente intratável.

3.3.2 Distribuições *a posteriori*

As distribuições *a posteriori* dos parâmetros A e $\mathbf{P}$ do modelo e da seqüência não observada $\mathbf{s}$ podem ser obtidas por simulação através do método *Gibbs sampling* com dados aumentados. Combinando a função de verossimilhança dada em (3.10) com a informação *a priori* sobre A e $\mathbf{P}$ e usando o teorema de Bayes, a distribuição *a posteriori* de A e $\mathbf{P}$ pode ser definida como

$$\pi\left(\mathbf{P}, A|\mathbf{y}, \mathbf{s}\right) \propto L(\mathbf{P}, A|\mathbf{y}, \mathbf{s})\pi\left(\mathbf{P}, A\right),$$

assumindo que A e $\mathbf{P}$ são independentes *a priori* temos

$$\pi\left(\mathbf{P}, A|\mathbf{y}, \mathbf{s}\right) \propto L(\mathbf{P}, A|\mathbf{y}, \mathbf{s})\pi\left(\mathbf{P}\right)\pi(A)$$

$$\begin{aligned}\pi\left(\mathbf{P},A|\mathbf{y},\mathbf{s}\right) &\propto \frac{1}{NM}\left(\prod_{k=1}^{N}\prod_{l=1}^{N}a_{kl}^{m_{kl}}\right)\left(\prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}}\right)\times \\ &\times\left(\prod_{k=1}^{N}\prod_{l=1}^{N}a_{kl}^{\beta_{kl}-1}\right)\left(\prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{\alpha_{ij}^{(k)}-1}\right) \\ &= \frac{1}{NM}\left(\prod_{k=1}^{N}\prod_{l=1}^{N}a_{kl}^{m_{kl}+\beta_{kl}-1}\right)\left(\prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}+\alpha_{ij}^{(k)}-1}\right). \quad (3.13)\end{aligned}$$

As distribuições *a posteriori* condicionais de $\mathbf{p}_i^{(k)}$ e $\mathbf{a}_k$, para $i = 1,2,...,M$ e $k = 1,2,...,N$, são, respectivamente,

$$\pi\left(\mathbf{p}_i^{(k)}|\mathbf{y},\mathbf{s},A,\mathbf{P}_{\left(\mathbf{p}_i^{(k)}\right)}\right)\propto\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}+\alpha_{ij}^{(k)}-1},$$

onde $\mathbf{P}_{\left(\mathbf{p}_i^{(k)}\right)} = \mathbf{P}\backslash\mathbf{p}_i^{(k)}$, denota $\mathbf{P}$ com a linha $\mathbf{p}_i^{(k)}$ removida e

$$\pi\left(\mathbf{a}_k|\mathbf{y},\mathbf{s},A_{(\mathbf{a}_k)},\mathbf{P}\right)\propto\prod_{l=1}^{N}a_{kl}^{m_{kl}+\beta_{kl}-1},$$

onde $A_{(\mathbf{a}_k)} = A\backslash\mathbf{a}_k$, denota A com a linha $\mathbf{a}_k$ removida, ou seja,

$$\mathbf{p}_i^{(k)}|\mathbf{y},\mathbf{s},A,\mathbf{P}_{\left(\mathbf{p}_i^{(k)}\right)}\sim\mathcal{D}(\mathbf{n}_i^{(k)}+\boldsymbol{\alpha}_i^{(k)}),$$

para $i = 1,2,...,M$ e $k = 1,2,...,N$ e

$$\mathbf{a}_k|\mathbf{y},\mathbf{s},A_{(\mathbf{a}_k)},\mathbf{P}\sim\mathcal{D}(\mathbf{m}_k+\boldsymbol{\beta}_k),$$

para $k = 1,2,...,N$ e onde

$$\begin{aligned}\mathbf{n}_i^{(k)} &= \{n_{ij}^k\},\ \text{para } j = 1,2,...,M \text{ e} \\ \mathbf{m}_k &= \{m_{kl}\},\ \text{para } l = 1,2,...,N.\end{aligned}$$

Estas distribuições formam as componentes da distribuição *a posteriori* de $\mathbf{P}$ e A, $\pi\left(\mathbf{P},A|\mathbf{y},\mathbf{s}\right)$.

Outra componente do *Gibbs sampling* é a distribuição da seqüência de estados ocultos, denominada $\pi(\mathbf{s}|\mathbf{y}, A, \mathbf{P})$. Um modo de simulação deste bloco é considerar T blocos univariados. Assim, a seqüência de estados ocultos é obtida através da simulação univariada e seqüencial de $\pi\left(s_t|\mathbf{s}_{(t)}, \mathbf{y}, A, \mathbf{P}\right)$, para $t = 1, 2, ..., T$ e onde $\mathbf{s}_{(t)} = \mathbf{s}\backslash s_t$ denota a seqüência $\mathbf{s}$ com o elemento t removido. A estrutura de dependência destas distribuições pode ser visualizada no grafo moral representado na Figura 3.3. Através do grafo, observamos que, para $t = 1, 2, ..., T$ e $s_t = 1, 2, ..., N$,

$$\begin{aligned}
\pi\left(s_t|\mathbf{s}_{(t)}, \mathbf{y}, A, \mathbf{P}\right) &= \pi\left(s_t|s_{t-1}, s_{t+1}, y_t, y_{t-1}, A, \mathbf{P}\right) \\
&= \frac{\Pr(S_t = s_t, S_{t+1} = s_{t+1}, Y_t = y_t|s_{t-1}, y_{t-1}, A, \mathbf{P})}{\Pr(S_{t+1} = s_{t+1}, Y_t = y_t|s_{t-1}, y_{t-1}, A, \mathbf{P})} \\
&= \frac{\left(\begin{array}{c}\Pr(S_t = s_t|s_{t-1}, y_{t-1}, A, \mathbf{P}) \Pr(Y_t = y_t|s_t, s_{t-1}, y_{t-1}, A, \mathbf{P})\times \\ \times \Pr(S_{t+1} = s_{t+1}|y_t, s_t, s_{t-1}, y_{t-1}, A, \mathbf{P})\end{array}\right)}{\sum_{k=1}^{N} \Pr(S_t = k, S_{t+1} = s_{t+1}, Y_t = y_t|s_{t-1}, y_{t-1}, A, \mathbf{P})},
\end{aligned}$$

pelas relações de independência condicional,

$$\pi\left(s_t|\mathbf{s}_{(t)}, \mathbf{y}, A, \mathbf{P}\right) = \frac{a_{s_{t-1}s_t} p^{(s_t)}_{y_{t-1}y_t} a_{s_t s_{t+1}}}{\sum_{k=1}^{N} a_{s_{t-1}k} p^{(k)}_{y_{t-1}y_t} a_{k s_{t+1}}}. \tag{3.14}$$

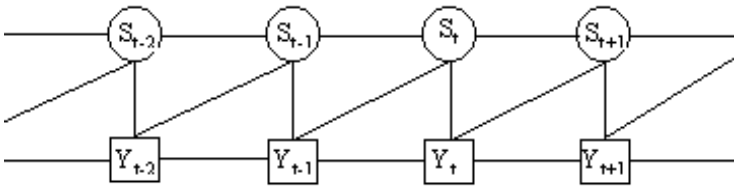


Figura 3.3: Grafo moral

Uma desvantagem do uso deste método para simulação de $\pi(\mathbf{s}|\mathbf{y}, A, \mathbf{P})$ é a dificuldade de convergência do algoritmo quando há um número grande de blocos componentes (Muri, 1997). Uma estratégia alternativa de simulação para a obtenção de realizações da distribuição $\pi(\mathbf{s}|\mathbf{y}, A, \mathbf{P})$ faz uso de uma outra propriedade de independência condicional. Seja $\mathbf{Y}^t = (Y_1, Y_2, ..., Y_t)$ a seqüência de estados observáveis até o tempo t. Através do

grafo direcional acíclico na Figura 3.2 observamos que S_t e Y_i $(i > t)$ são condicionalmente independentes dado S_{t+1} e $\mathbf{Y}^t$. Desta maneira, excluindo $\mathbf{P}$ e A da notação por simplicidade, temos para $t = 1, 2, ..., T-1$.

$$\begin{aligned}
\Pr(S_t &= s_t|s_{t+1}, \mathbf{y}) = \Pr(S_t = s_t|s_{t+1}, \mathbf{y}^t) \\
&= \frac{\Pr(S_t = s_t, S_{t+1} = s_{t+1}|\mathbf{y}^t)}{\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)} \\
&= \frac{\Pr(S_t = s_t|\mathbf{y}^t)\Pr(S_{t+1} = s_{t+1}|s_t, \mathbf{y}^t)}{\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)} \\
&= \frac{a_{s_t s_{t+1}}\Pr(S_t = s_t|\mathbf{y}^t)}{\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)}.
\end{aligned} \tag{3.15}$$

Desenvolvendo $\Pr(S_t = s_t|\mathbf{y}^t)$ e $\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)$ temos, respectivamente,

$$\begin{aligned}
\Pr(S_t &= s_t|\mathbf{y}^t) = \Pr(S_t = s_t|y_t, \mathbf{y}^{t-1}) \\
&= \frac{\Pr(S_t = s_t, Y_t = y_t|\mathbf{y}^{t-1})}{\Pr(Y_t = y_t|\mathbf{y}^{t-1})} \\
&\propto \Pr(S_t = s_t, Y_t = y_t|\mathbf{y}^{t-1}) \\
&= \sum_{l=1}^{N}\Pr(S_t = s_t, Y_t = y_t, S_{t-1} = l|\mathbf{y}^{t-1}) \\
&= \sum_{l=1}^{N}\left(\begin{array}{c}\Pr(S_{t-1} = l|\mathbf{y}^{t-1})\Pr(S_t = s_t|S_{t-1} = l, \mathbf{y}^{t-1})\times \\ \times \Pr(Y_t = y_t|S_{t-1} = l, s_t, \mathbf{y}^{t-1})\end{array}\right),
\end{aligned}$$

pelas relações de independência condicional, temos

$$\begin{aligned}
\Pr(S_t &= s_t|\mathbf{y}^t) \propto \Pr(Y_t = y_t|s_t, y_{t-1})\sum_{l=1}^{N}\Pr(S_{t-1} = l|\mathbf{y}^{t-1})\Pr(S_t = s_t|S_{t-1} = l) \\
&= p_{y_{t-1}y_t}^{(s_t)}\sum_{l=1}^{N}a_{ls_t}\Pr(S_{t-1} = l|\mathbf{y}^{t-1}) \text{ e}
\end{aligned} \tag{3.16}$$

$$\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t) = \sum_{k=1}^{N}\Pr(S_{t+1} = s_{t+1}, S_t = k|\mathbf{y}^t)$$

$$\begin{aligned}\Pr(S_{t+1} &= s_{t+1}|\mathbf{y}^t) = \sum_{k=1}^{N} \Pr(S_t = k|\mathbf{y}^t) \Pr(S_{t+1} = s_{t+1}|S_t = k, \mathbf{y}^t) \\ &= \sum_{k=1}^{N} a_{k s_{t+1}} \Pr(S_t = k|\mathbf{y}^t). \end{aligned} \tag{3.17}$$

A equação (3.16) fornece um esquema (*forward*) iterativo para determinar valores de $\Pr(S_t = s_t|\mathbf{y}^t)$, para $t = 1, 2, ..., T$. A distribuição inicial para as iterações é fornecida pela distribuição discreta uniforme sobre (Y_1, S_1). Valores de $\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)$ e $\Pr(S_t = s_t|s_{t+1}, \mathbf{y})$ podem ser calculados através das equações (3.17) e (3.15), respectivamente.

Uma realização do processo aleatório $\mathbf{S}$ pode, então, ser simulada. Primeiro, um valor de s_T é obtido pelo uso da distribuição $\Pr(S_T = s_T|\mathbf{y}^T)$ e, assim, os valores restantes são obtidos, através de um procedimento *backward*, usando a equação (3.15), para $t = T - 1, T - 2, ..., 1$.

A performance do *Gibbs sampling* e, em particular, sua convergência, pode ser verificada através do uso de uma variedade de gráficos e diagnósticos numéricos.

Batistela (2003) estudaram o desempenho dos estimadores bayesianos para o HMM$(1, 1)$ em diversas situações e definiram em que condições as estimativas são mais precisas e corretas.

3.4 HMM com dependência de Markov de primeira ordem entre os estados ocultos e de segunda ordem entre os estados observáveis, HMM$(1, 2)$

Um HMM mais complexo do que os apresentados nas seções anteriores, é o modelo cuja seqüência de variáveis observáveis é considerada uma cadeia de Markov de segunda ordem, ou seja, assumimos que transições entre os estados observáveis (resultados) $Y_{t-1} \rightarrow Y_t$ segue um processo estocástico, cuja escolha da matriz de transição é determinada pelo estado oculto S_t e pelo estado observável Y_{t-2}.

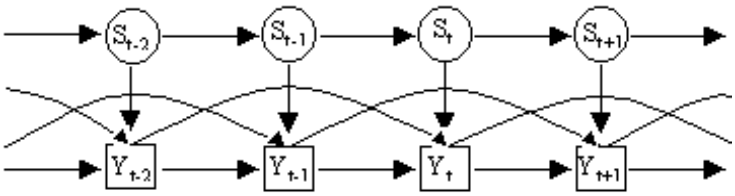


Figura 3.4: Grafo Direcional Acíclico do HMM$(1, 2)$

Uma maneira de entender a estrutura de dependência do modelo é através do grafo direcional acíclico, representado na Figura 3.4, no qual podemos observar que o modelo obedece as seguintes relações de independência condicional,

$$S_t \perp \{S_1, Y_1, ..., S_{t-2}, Y_{t-2}, Y_{t-1}\} | S_{t-1},\ 2 \leq t \leq T \text{ e} \tag{3.18}$$

$$Y_t \perp \{S_1, Y_1, ..., S_{t-3}, Y_{t-3}, S_{t-2}, S_{t-1}\} | Y_{t-2}, Y_{t-1} \text{ e } S_t,\ 3 \leq t \leq T\ . \tag{3.19}$$

Assumindo que S_1, Y_1, S_2 e Y_2 seguem distribuições uniformes discretas independentes, este HMM é especificado pelos seguintes elementos:

$A = \{a_{kl}\}$, representa a matriz com as probabilidades de transição entre os estados ocultos, onde

$$a_{kl} = \Pr(S_{t+1} = l | S_t = k, A) \tag{3.20}$$

é a probabilidade de estar no estado l no tempo $t+1$ dado que no instante t o estado era k, para $1 \leq k, l \leq N$ e;

$\mathbf{P} = \left\{P^{(1)}, P^{(2)}, ..., P^{(N)}\right\}$, onde $P^{(k)} = \left\{p_{hij}^{(k)}\right\}$, representa a matriz com as probabilidades de transição entre os estados observáveis (resultados) associada ao k-ésimo estado oculto, e

$$p_{hij}^{(k)} = \Pr(Y_t = j | S_t = k, Y_{t-1} = i, Y_{t-2} = h, \mathbf{P}) \tag{3.21}$$

é a probabilidade de observar o resultado j no instante t dado que a cadeia oculta está no estado k, no instante anterior $t-1$ foi observado o resultado i e no instante $t-2$ foi observado o resultado h, para $k = 1, 2, ..., N$ e $1 \leq h, i, j \leq M$.

Assim, a função de verossimilhança para os parâmetros $\mathbf{P}$ e A do modelo, dada a seqüência observada $\mathbf{y}$ e a não observada $\mathbf{s}$, é

$$L(\mathbf{P}, A|\mathbf{y}, \mathbf{s}) = \Pr(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s}|A, \mathbf{P}),$$

que, pelas relações de independência condicional e suposições do modelo, pode ser escrita como

$$\begin{aligned}
L(\mathbf{P}, A|\mathbf{y}, \mathbf{s}) &= \left(\prod_{t=1}^{2} \Pr(S_t = s_t|A, \mathbf{P}) \Pr(Y_t = y_t|A, \mathbf{P})\right) \times \\
&\quad \times \left(\prod_{t=3}^{T} \Pr(S_t = s_t|s_{t-1}, A, \mathbf{P}) \Pr(Y_t = y_t|y_{t-2}, y_{t-1}, s_t, A, \mathbf{P})\right) \\
&= \frac{1}{(NM)^2} \prod_{t=3}^{T} a_{s_{t-1}s_t} p^{(s_t)}_{y_{t-2}y_{t-1}y_t} \\
&= \frac{1}{(NM)^2} \left(\prod_{t=3}^{T} a_{s_{t-1}s_t}\right) \left(\prod_{t=3}^{T} p^{(s_t)}_{y_{t-2}y_{t-1}y_t}\right) \\
&= \frac{1}{(NM)^2} \left(\prod_{k=1}^{N}\prod_{l=1}^{N} a_{kl}^{m_{kl}}\right) \left(\prod_{k=1}^{N}\prod_{h=1}^{M}\prod_{i=1}^{M}\prod_{j=1}^{M} \left(p^{(k)}_{hij}\right)^{n^{(k)}_{hij}}\right), \qquad (3.22)
\end{aligned}$$

onde

$$\begin{aligned}
m_{kl} &= \sum_{t=2}^{T-1} I\left(s_t = k, s_{t+1} = l\right) \text{ e} \\
n^{(k)}_{hij} &= \sum_{t=3}^{T} I\left(y_{t-2} = h, y_{t-1} = i, y_t = j, s_t = k\right),
\end{aligned}$$

onde $I(A) = \begin{cases} 1 \text{ se } A \text{ é verdade} \\ 0 \text{ caso contrário} \end{cases}$.

No HMM$(1, 2)$, as inferências sobre A e $\mathbf{P}$ também podem ser realizadas através de métodos bayesianos com simulação da distribuição *a posteriori* de A e $\mathbf{P}$ usando o método Monte Carlo em cadeias de Markov (MCMC), assim como a estimação dos estados ocultos, que são tratados como variáveis não observáveis e são simuladas da distribuição condicional.

3.4.1 Distribuições *a priori*

Seja $\mathbf{p}_{hi}^{(k)} = \{p_{hij}^{(k)}\}$ um vetor da matriz tripla de transição entre os estados observáveis associada ao k-ésimo estado oculto e $\mathbf{a}_k = \{a_{kl}\}$ uma linha da matriz de transição entre os estados ocultos A, para $1 \leq h, i, j \leq M$, $1 \leq k \leq N$ e todos supostamente independentes. Dada a forma multinomial da função de verossimilhança, uma distribuição *a priori* conjugada para o vetor $M-$dimensional $\mathbf{p}_{hi}^{(k)}$ e para o vetor $N-$dimensional $\mathbf{a}_k$ é uma distribuição Dirichlet definida no simplex com densidade, respectivamente,

$$\pi\left(\mathbf{p}_{hi}^{(k)}\right) \propto \prod_{j=1}^{M} \left(p_{hij}^{(k)}\right)^{\alpha_{hij}^{(k)}-1},$$

para $0 < p_{hij}^{(k)} < 1$, $\sum_{j=1}^{M} p_{hij}^{(k)} = 1$, $k = 1, 2, ..., N$, $1 \leq h, i, j \leq M$ e onde $\boldsymbol{\alpha}_{hi}^{(k)} = \{\alpha_{hij}^{(k)}\}$ são os parâmetros positivos da distribuição, e

$$\pi\left(\mathbf{a}_k\right) \propto \prod_{l=1}^{N} a_{kl}^{\beta_{kl}-1},$$

para $0 < a_{kl} < 1$, $\sum_{l=1}^{n} a_{kl} = 1$, $k = 1, 2, ..., N$ e onde $\boldsymbol{\beta}_k = \{\beta_{kl}\}$ são os parâmetros positivos da distribuição.

Assim, a distribuição *a priori* para A e $\mathbf{P}$ são, respectivamente, definidas como

$$\begin{aligned} \pi\left(A\right) &\propto \prod_{k=1}^{N} \pi\left(\mathbf{a}_k\right) \\ &= \prod_{k=1}^{N} \prod_{l=1}^{N} a_{kl}^{\beta_{kl}-1} \text{ e} \end{aligned} \tag{3.23}$$

$$\begin{aligned} \pi\left(\mathbf{P}\right) &\propto \prod_{k=1}^{N} \prod_{h=1}^{M} \prod_{i=1}^{M} \pi\left(\mathbf{p}_{hi}^{(k)}\right) \\ &= \prod_{k=1}^{N} \prod_{h=1}^{M} \prod_{i=1}^{M} \prod_{j=1}^{M} \left(p_{hij}^{(k)}\right)^{\alpha_{hij}^{(k)}-1}. \end{aligned} \tag{3.24}$$

3.4.2 Distribuições *a posteriori*

As distribuições *a posteriori* dos parâmetros A e $\mathbf{P}$ do modelo e da seqüência não observável $\mathbf{s}$ podem ser obtidas e simuladas através do método *Gibbs sampling* com dados aumentados. Combinando a função de verossimilhança dada em (3.22) com a informação *a priori* sobre A e $\mathbf{P}$ e usando o teorema de Bayes, a distribuição *a posteriori* de A e $\mathbf{P}$ pode ser definida como

$$\pi\left(\mathbf{P}, A|\mathbf{y}, \mathbf{s}\right) \propto L(\mathbf{P}, A|\mathbf{y}, \mathbf{s})\pi\left(\mathbf{P}, A\right),$$

assumindo que A e $\mathbf{P}$ são independentes *a priori* temos

$$\begin{aligned}
\pi\left(\mathbf{P}, A|\mathbf{y}, \mathbf{s}\right) &\propto L(\mathbf{P}, A|\mathbf{y}, \mathbf{s})\pi\left(\mathbf{P}\right)\pi(A) \\
&= \frac{1}{(NM)^2}\left(\prod_{k=1}^{N}\prod_{l=1}^{N} a_{kl}^{m_{kl}}\right)\left(\prod_{k=1}^{N}\prod_{h=1}^{M}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{hij}^{(k)}\right)^{n_{hij}^{(k)}}\right) \times \\
&\quad \times \left(\prod_{k=1}^{N}\prod_{l=1}^{N} a_{kl}^{\beta_{kl}-1}\right)\left(\prod_{k=1}^{N}\prod_{h=1}^{M}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{hij}^{(k)}\right)^{\alpha_{hij}^{(k)}-1}\right) \\
&= \frac{1}{(NM)^2}\left(\prod_{k=1}^{N}\prod_{l=1}^{N} a_{kl}^{m_{kl}+\beta_{kl}-1}\right)\left(\prod_{k=1}^{N}\prod_{h=1}^{M}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{hij}^{(k)}\right)^{n_{hij}^{(k)}+\alpha_{hij}^{(k)}-1}\right).
\end{aligned} \tag{3.25}$$

As distribuições condicionais de $\mathbf{p}_{hi}^{(k)}$ e $\mathbf{a}_k$, para $1 \leq h, i \leq M$ e $k = 1, 2, ..., N$, são, respectivamente,

$$\pi\left(\mathbf{p}_{hi}^{(k)}|\mathbf{y}, \mathbf{s}, A, \mathbf{P}_{\left(\mathbf{p}_{hi}^{(k)}\right)}\right) \propto \prod_{j=1}^{M}\left(p_{hij}^{(k)}\right)^{n_{hij}^{(k)}+\alpha_{hij}^{(k)}-1},$$

onde $\mathbf{P}_{\left(\mathbf{p}_{hi}^{(k)}\right)} = \mathbf{P}\backslash\mathbf{p}_{hi}^{(k)}$, denota $\mathbf{P}$ com a linha $\mathbf{p}_{hi}^{(k)}$ removida e

$$\pi\left(\mathbf{a}_k|\mathbf{y}, \mathbf{s}, A_{(\mathbf{a}_k)}, \mathbf{P}\right) \propto \prod_{l=1}^{N} a_{kl}^{m_{kl}+\beta_{kl}-1},$$

onde $A_{(\mathbf{a}_k)} = A\backslash\mathbf{a}_k$, denota A com a linha $\mathbf{a}_k$ removida, ou seja,

$$\mathbf{p}_{hi}^{(k)}|\mathbf{y},\mathbf{s},A,\mathbf{P}_{\left(\mathbf{p}_{hi}^{(k)}\right)} \sim \mathcal{D}(\mathbf{n}_{hi}^{(k)}+\boldsymbol{\alpha}_{hi}^{(k)}),$$

para $1 \leq h,i \leq M$ e $k = 1,2,...,N$ e

$$\mathbf{a}_k|\mathbf{y},\mathbf{s},A_{(\mathbf{a}_k)},\mathbf{P} \sim \mathcal{D}(\mathbf{m}_k+\boldsymbol{\beta}_k),$$

para $k = 1,2,...,N$ e onde

$$\begin{aligned}\mathbf{n}_{hi}^{(k)} &= \{n_{hij}^{k}\},\ \text{para } j = 1,2,...,M \text{ e}\\ \mathbf{m}_k &= \{m_{kl}\},\ \text{para } l = 1,2,...,N.\end{aligned}$$

Outra componente do *Gibbs sampling* é a distribuição da seqüência de estados ocultos, $\pi(\mathbf{s}|\mathbf{y},A,\mathbf{P})$.

Um método de simulação deste bloco é considerá-lo como T blocos univariados. Assim, a seqüência de estados ocultos é obtida através da simulação univariada e seqüencial de $\pi\left(s_t|\mathbf{s}_{(t)},\mathbf{y},A,\mathbf{P}\right)$, para $t = 1,2,...,T$ e onde $\mathbf{s}_{(t)} = \mathbf{s}\backslash s_t$ denota a seqüência $\mathbf{s}$ com o elemento t removido. A estrutura de dependência destas distribuições pode ser visualizada no grafo moral representado na Figura 3.5. Através do grafo, observamos que, para $t = 1,2,...,T$ e $s_t = 1,2,...,N$,

$$\begin{aligned}\pi\left(s_t|\mathbf{s}_{(t)},\mathbf{y},A,\mathbf{P}\right) &= \pi(s_t|s_{t-1},s_{t+1},y_t,y_{t-1},y_{t-2},A,\mathbf{P})\\ &= \frac{\Pr(S_t=s_t,S_{t+1}=s_{t+1},Y_t=y_t|s_{t-1},y_{t-1},y_{t-2},A,\mathbf{P})}{\Pr(S_{t+1}=s_{t+1},Y_t=y_t|s_{t-1},y_{t-1},y_{t-2},A,\mathbf{P})},\end{aligned}$$

pelas relações de independência condicional,

$$\pi\left(s_t|\mathbf{s}_{(t)},\mathbf{y},A,\mathbf{P}\right) = \frac{a_{s_{t-1}s_t}p_{y_{t-2}y_{t-1}y_t}^{(s_t)}a_{s_ts_{t+1}}}{\sum_{k=1}^{N}a_{s_{t-1}k}p_{y_{t-2}y_{t-1}y_t}^{(k)}a_{ks_{t+1}}}. \tag{3.26}$$

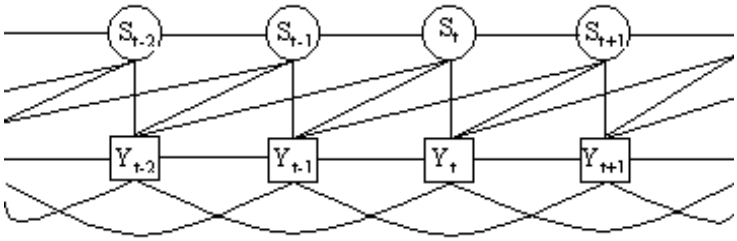


Figura 3.5: Grafo moral

Uma estratégia alternativa de simulação para a obtenção de realizações da distribuição $\pi(\mathbf{s}|\mathbf{y}, A, \mathbf{P})$ faz uso de uma propriedade de independência condicional diferente. Seja $\mathbf{Y}^t = (Y_1, Y_2, ..., Y_t)$ a seqüência de estados observáveis até o tempo t. Através do grafo direcional acíclico na Figura 3.4 observamos que S_t e Y_i $(i > t)$ são condicionalmente independentes dado S_{t+1} e $\mathbf{Y}^t$. Desta maneira, excluindo $\mathbf{P}$ e A da notação por simplicidade, temos para $t = 1, 2, ..., T-1$.

$$\begin{aligned} \Pr(S_t &= s_t|s_{t+1}, \mathbf{y}) = \Pr(S_t = s_t|s_{t+1}, \mathbf{y}^t) \\ &= \frac{a_{s_t s_{t+1}} \Pr(S_t = s_t|\mathbf{y}^t)}{\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)}. \end{aligned} \tag{3.27}$$

Desenvolvendo $\Pr(S_t = s_t|\mathbf{y}^t)$ e $\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)$ temos, respectivamente,

$$\begin{aligned} \Pr(S_t &= s_t|\mathbf{y}^t) = \Pr(S_t = s_t|y_t, \mathbf{y}^{t-1}) \\ &\propto p^{(s_t)}_{y_{t-2}y_{t-1}y_t} \sum_{l=1}^{N} a_{ls_t} \Pr(S_{t-1} = l|\mathbf{y}^{t-1}) \text{ e} \end{aligned} \tag{3.28}$$

$$\begin{aligned} \Pr(S_{t+1} &= s_{t+1}|\mathbf{y}^t) = \sum_{k=1}^{N} \Pr(S_{t+1} = s_{t+1}, S_t = k|\mathbf{y}^t) \\ &= \sum_{k=1}^{N} a_{ks_{t+1}} \Pr(S_t = k|\mathbf{y}^t). \end{aligned} \tag{3.29}$$

A equação (3.28) fornece um esquema (*forward*) iterativo para determinar $\Pr(S_t = s_t|\mathbf{y}^t)$, para $t = 1, 2, ..., T$. A distribuição inicial para as iterações é fornecida pela distribuição discreta uniforme sobre (Y_1, S_1) e (Y_2, S_2). Valores de $\Pr(S_{t+1} = s_{t+1}|\mathbf{y}^t)$ e

$\Pr(S_t = s_t|s_{t+1}, \mathbf{y})$ podem ser calculados através das equações (3.29) e (3.27), respectivamente.

Uma realização do processo aleatório $\mathbf{S}$ pode agora ser simulada. Primeiro, um valor de s_T é obtido pelo uso da distribuição $\Pr(S_T = s_T|\mathbf{y}^T)$ e, assim, os valores restantes são obtidos, através de um procedimento *backward*, usando a equação (3.27), para $t = T-1, T-2, ..., 1$.

A performance do método *Gibbs sampling* e, em particular, sua convergência, pode ser verificada através do uso de uma variedade de gráficos e diagnósticos numéricos.

3.5 HMM com dependência de Markov de segunda ordem entre os estados ocultos e de primeira ordem entre os estados observáveis, HMM$(2, 1)$

Nas seções anteriores, apresentamos três diferentes HMMs com dependência de Markov de primeira ordem na cadeia de Markov oculta $\mathbf{S}$. Nesta seção, introduzimos um HMM cuja seqüência oculta é considerada uma cadeia de Markov de segunda ordem, ou seja, as transições entre os estados ocultos S_t seguem uma cadeia de Markov, cuja escolha da matriz de transição é determinada pelos estados ocultos S_{t-2} e S_{t-1}. No HMM$(2, 1)$, a seqüência de estados observáveis $\mathbf{Y}$ é uma cadeia de Markov de primeira ordem.

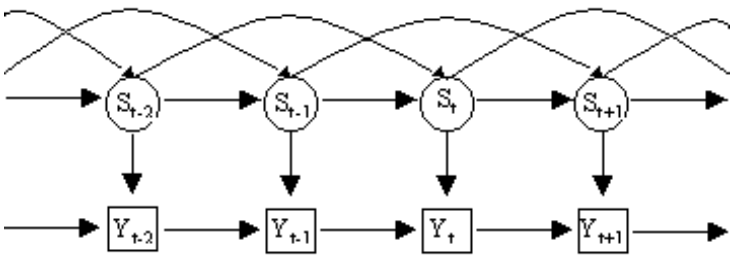


Figura 3.6: Grafo Direcional Acíclico do HMM$(2, 1)$

A estrutura de dependência deste modelo é mais facilmente compreendida através do grafo acíclico direcional representado na Figura 3.6, no qual podemos observar que o modelo obedece as seguintes relações de independência condicional,

$$S_t \perp \{S_1, Y_1, ..., S_{t-3}, Y_{t-3}, Y_{t-2}, Y_{t-1}\}|S_{t-1} \text{ e } S_{t-2},\ 3 \leq t \leq T \text{ e} \qquad (3.30)$$

$$Y_t \perp \{S_1, Y_1, ..., S_{t-2}, Y_{t-2}, S_{t-1}\} | S_t \text{ e } Y_{t-1},\ 2 \leq t \leq T\ . \tag{3.31}$$

Assumindo que S_1, Y_1, S_2 e Y_2 seguem distribuições uniformes discretas independentes, o HMM$(2,1)$ é especificado pelos seguintes elementos:

$A = \{a_{hkl}\}$, representa a matriz com as probabilidades de transição entre os estados ocultos, onde

$$a_{hkl} = \Pr(S_{t+1} = l | S_t = k, S_{t-1} = h, A), \tag{3.32}$$

para $1 \leq h, k, l \leq N$ e;

$\mathbf{P} = \left\{P^{(1)}, P^{(2)}, ..., P^{(N)}\right\}$, onde $P^{(k)} = \left\{p_{ij}^{(k)}\right\}$, representa a matriz com as probabilidades de transição entre os estados observáveis associada ao k-ésimo estado oculto, e

$$p_{ij}^{(k)} = \Pr(Y_t = j | S_t = k, Y_{t-1} = i, \mathbf{P}) \tag{3.33}$$

é a probabilidade de observar o resultado j em t dado que a cadeia oculta está no estado k e no instante anterior $t-1$ foi observado o resultado i, para $k = 1, 2, ..., N$ e $1 \leq i, j \leq M$.

Assim, a função de verossimilhança para os parâmetros $\mathbf{P}$ e A do modelo, dada a seqüência observada $\mathbf{y}$ e a não observada $\mathbf{s}$, é

$$L(\mathbf{P}, A | \mathbf{y}, \mathbf{s}) = \Pr(\mathbf{Y} = \mathbf{y}, \mathbf{S} = \mathbf{s} | A, \mathbf{P}),$$

que, pelas relações de independência condicional e suposições do modelo, pode ser escrita como

$$\begin{aligned} L(\mathbf{P}, A | \mathbf{y}, \mathbf{s}) &= \left(\prod_{t=1}^{2} \Pr(S_t = s_t | A, \mathbf{P}) \Pr(Y_t = y_t | A, \mathbf{P})\right) \times \\ &\quad \times \left(\prod_{t=3}^{T} \Pr(S_t = s_t | s_{t-1}, s_{t-2}, A, \mathbf{P}) \Pr(Y_t = y_t | y_{t-1}, s_t, A, \mathbf{P})\right) \\ &= \frac{1}{(NM)^2} \prod_{t=3}^{T} a_{s_{t-2} s_{t-1} s_t} p_{y_{t-1} y_t}^{(s_t)} \end{aligned}$$

$$
\begin{aligned}
L(\mathbf{P}, A|\mathbf{y}, \mathbf{s}) &= \frac{1}{(NM)^2}\left(\prod_{t=3}^{T} a_{s_{t-2}s_{t-1}s_t}\right)\left(\prod_{t=3}^{T} p_{y_{t-1}y_t}^{(s_t)}\right) \\
&= \frac{1}{(NM)^2}\left(\prod_{h=1}^{N}\prod_{k=1}^{N}\prod_{l=1}^{N} a_{hkl}^{m_{hkl}}\right)\left(\prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}}\right), \qquad (3.34)
\end{aligned}
$$

onde

$$
\begin{aligned}
m_{hkl} &= \sum_{t=2}^{T-1} I\left(s_{t-1}=h, s_t=k, s_{t+1}=l\right) \text{ e} \\
n_{ij}^{(k)} &= \sum_{t=3}^{T} I\left(y_{t-1}=i, y_t=j, s_t=k\right),
\end{aligned}
$$

onde $I(A) = \begin{cases} 1 \text{ se } A \text{ é verdade} \\ 0 \text{ caso contrário} \end{cases}$.

No HMM$(2,1)$, as inferências sobre A e $\mathbf{P}$ podem ser realizadas através de métodos bayesianos com simulação da distribuição *a posteriori* de A e $\mathbf{P}$ usando o método Monte Carlo em cadeias de Markov (MCMC), assim como a estimação dos estados ocultos, que são tratados como variáveis não observáveis e são simuladas da distribuição condicional.

3.5.1 Distribuições *a priori*

Considere $\mathbf{p}_i = \{p_{ij}\}$ uma linha de uma das matrizes de transição dos estados observáveis, para $j = 1, 2, ..., M$. Dada a forma multinomial da função de verossimilhança, uma distribuição *a priori* conjugada para o vetor $M-$dimensional $\mathbf{p}_i$ é a distribuição Dirichlet definida no simplex, com densidade

$$
\pi\left(\mathbf{p}_i\right) \propto \prod_{j=1}^{M} p_{ij}^{\alpha_{ij}-1},
$$

onde $0 < p_{ij} < 1$, $j = 1, 2, ..., M$, $\sum_{j=1}^{M} p_{ij} = 1$ e $\boldsymbol{\alpha}_i = \{\alpha_{ij}\}$ são os parâmetros positivos da distribuição.

Supondo que as M linhas das N matrizes de transição entre os estados observáveis

são independentes *a priori* e que

$$\mathbf{p}_i^{(k)} = \{p_{ij}^{(k)}\} \sim \mathcal{D}(\boldsymbol{\alpha}_i^{(k)}),$$

para $k = 1, 2, ..., N$ e $i = 1, 2, ..., M$, temos que a distribuição *a priori* para $\mathbf{P}$ tem densidade

$$\begin{aligned} \pi(\mathbf{P}) &\propto \prod_{k=1}^{N}\prod_{i=1}^{M} \pi\left(\mathbf{p}_i^{(k)}\right) \\ &= \prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M} \left(p_{ij}^{(k)}\right)^{\alpha_{ij}^{(k)}-1}. \end{aligned} \tag{3.35}$$

Similarmente, usamos a distribuição Dirichlet N-dimensional como a distribuição *a priori* para os vetores $\mathbf{a}_{hk} = \{a_{hkl}\}$ da matriz de transição entre os estados ocultos. Assumindo que os vetores são independentes *a priori* e que

$$\mathbf{a}_{hk} = \{a_{hkl}\} \sim \mathcal{D}(\boldsymbol{\beta}_{hk}),$$

para $1 \leq h, k, l \leq N$ e onde $\boldsymbol{\beta}_{hk} = \{\beta_{hkl}\}$ são os parâmetros positivos da distribuição, temos que a distribuição *a priori* para A tem densidade

$$\begin{aligned} \pi(A) &\propto \prod_{h=1}^{N}\prod_{k=1}^{N} \pi(\mathbf{a}_{hk}) \\ &= \prod_{h=1}^{N}\prod_{k=1}^{N}\prod_{l=1}^{N} a_{hkl}^{\beta_{hkl}-1}. \end{aligned} \tag{3.36}$$

3.5.2 Distribuições *a posteriori*

As distribuições *a posteriori* dos parâmetros A e $\mathbf{P}$ do modelo e da seqüência não observável $\mathbf{s}$ podem ser obtidas e simuladas através do método *Gibbs sampling* com dados aumentados. Combinando a função de verossimilhança dada em (3.34) com a informação *a priori* sobre A e $\mathbf{P}$ e usando o teorema de Bayes, a distribuição *a posteriori* de A e $\mathbf{P}$ pode ser definida como

$$\pi\left(\mathbf{P}, A|\mathbf{y}, \mathbf{s}\right) \propto L(\mathbf{P}, A|\mathbf{y}, \mathbf{s})\pi\left(\mathbf{P}, A\right),$$

assumindo que A e $\mathbf{P}$ são independentes *a priori* temos

$$\begin{aligned}
\pi\left(\mathbf{P}, A|\mathbf{y}, \mathbf{s}\right) &\propto L(\mathbf{P}, A|\mathbf{y}, \mathbf{s})\pi\left(\mathbf{P}\right)\pi(A) \\
&= \frac{1}{(NM)^2}\left(\prod_{h=1}^{N}\prod_{k=1}^{N}\prod_{l=1}^{N} a_{hkl}^{m_{hkl}}\right)\left(\prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}}\right)\times \\
&\quad \times\left(\prod_{h=1}^{N}\prod_{k=1}^{N}\prod_{l=1}^{N} a_{hkl}^{\beta_{hkl}-1}\right)\left(\prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{\alpha_{ij}^{(k)}-1}\right) \\
&= \frac{1}{(NM)^2}\left(\prod_{h=1}^{N}\prod_{k=1}^{N}\prod_{l=1}^{N} a_{hkl}^{m_{hkl}+\beta_{hkl}-1}\right)\left(\prod_{k=1}^{N}\prod_{i=1}^{M}\prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}+\alpha_{ij}^{(k)}-1}\right).
\end{aligned} \tag{3.37}$$

As distribuições *a posteriori* condicionais de $\mathbf{p}_i^{(k)}$ e $\mathbf{a}_{hk}$, para $i = 1, 2, ..., M$ e $1 \leq h, k \leq N$, são, respectivamente,

$$\pi\left(\mathbf{p}_i^{(k)}|\mathbf{y}, \mathbf{s}, A, \mathbf{P}_{\left(\mathbf{p}_i^{(k)}\right)}\right) \propto \prod_{j=1}^{M}\left(p_{ij}^{(k)}\right)^{n_{ij}^{(k)}+\alpha_{ij}^{(k)}-1},$$

onde $\mathbf{P}_{\left(\mathbf{p}_i^{(k)}\right)} = \mathbf{P}\backslash\mathbf{p}_i^{(k)}$, denota $\mathbf{P}$ com a linha $\mathbf{p}_i^{(k)}$ removida e

$$\pi\left(\mathbf{a}_{hk}|\mathbf{y}, \mathbf{s}, A_{(\mathbf{a}_{hk})}, \mathbf{P}\right) \propto \prod_{l=1}^{N} a_{hkl}^{m_{hkl}+\beta_{hkl}-1},$$

onde $A_{(\mathbf{a}_{hk})} = A\backslash\mathbf{a}_{hk}$, denota A com a linha $\mathbf{a}_{hk}$ removida, ou seja,

$$\mathbf{p}_i^{(k)}|\mathbf{y}, \mathbf{s}, A, \mathbf{P}_{\left(\mathbf{p}_i^{(k)}\right)} \sim \mathcal{D}(\mathbf{n}_i^{(k)} + \boldsymbol{\alpha}_i^{(k)}),$$

para $i = 1, 2, ..., M$ e $k = 1, 2, ..., N$ e

$$\mathbf{a}_{hk}|\mathbf{y}, \mathbf{s}, A_{(\mathbf{a}_{hk})}, \mathbf{P} \sim \mathcal{D}(\mathbf{m}_{hk} + \boldsymbol{\beta}_{hk}),$$

para $1 \leq h, k \leq N$ e onde

$$\begin{aligned}\mathbf{n}_i^{(k)} &= \{n_{ij}^k\}, \text{ para } j = 1, 2, ..., M \text{ e} \\ \mathbf{m}_{hk} &= \{m_{hkl}\}, \text{ para } l = 1, 2, ..., N.\end{aligned}$$

Outra componente do método *Gibbs sampling* é a distribuição da seqüência de estados ocultos, $\pi(\mathbf{s}|\mathbf{y}, A, \mathbf{P})$.

Um método de simulação deste bloco é considerá-lo como T blocos univariados. Assim, a seqüência de estados ocultos é obtida através da simulação univariada e seqüencial de $\pi\left(s_t|\mathbf{s}_{(t)}, \mathbf{y}, A, \mathbf{P}\right)$, para $t = 1, 2, ..., T$ e onde $\mathbf{s}_{(t)} = \mathbf{s}\backslash s_t$ denota a seqüência $\mathbf{s}$ com o elemento t removido. A estrutura de dependência destas distribuições pode ser visualizada no grafo moral representado na Figura 3.7. Através do grafo, observamos que, para $t = 1, 2, ..., T$ e $s_t = 1, 2, ..., N$,

$$\begin{aligned}\pi\left(s_t|\mathbf{s}_{(t)}, \mathbf{y}, A, \mathbf{P}\right) &= \pi(s_t|s_{t-2}, s_{t-1}, s_{t+1}, s_{t+2}, y_t, y_{t-1}, A, \mathbf{P}) \\ &= \frac{\Pr(S_t = s_t, S_{t+1} = s_{t+1}, S_{t+2} = s_{t+2}, Y_t = y_t|s_{t-2}, s_{t-1}, y_{t-1}, A, \mathbf{P})}{\Pr(S_{t+1} = s_{t+1}, S_{t+2} = s_{t+2}, Y_t = y_t|s_{t-2}, s_{t-1}, y_{t-1}, A, \mathbf{P})} \\ &= \frac{\left(\begin{array}{c}\Pr(S_t = s_t|s_{t-2}, s_{t-1}, y_{t-1}, A, \mathbf{P})\times \\ \times \Pr(Y_t = y_t|s_t, s_{t-2}, s_{t-1}, y_{t-1}, A, \mathbf{P})\times \\ \times \Pr(S_{t+1} = s_{t+1}|y_t, s_t, s_{t-2}, s_{t-1}, y_{t-1}, A, \mathbf{P})\times \\ \times \Pr(S_{t+2} = s_{t+2}|s_{t+1}, y_t, s_t, s_{t-2}, s_{t-1}, y_{t-1}, A, \mathbf{P})\end{array}\right)}{\sum_{k=1}^{N} PP_k},\end{aligned}$$

onde

$$PP_k = \Pr(S_t = k, S_{t+1} = s_{t+1}, S_{t+2} = s_{t+2}, Y_t = y_t|s_{t-2}, s_{t-1}, y_{t-1}, A, \mathbf{P}),$$

e pelas relações de independência condicional,

$$\pi\left(s_t|\mathbf{s}_{(t)}, \mathbf{y}, A, \mathbf{P}\right) = \frac{a_{s_{t-2}s_{t-1}s_t} p_{y_{t-1}y_t}^{(s_t)} a_{s_{t-1}s_t s_{t+1}} a_{s_t s_{t+1} s_{t+2}}}{\sum_{k=1}^{N} a_{s_{t-2}s_{t-1}k} p_{y_{t-1}y_t}^{(k)} a_{s_{t-1}k s_{t+1}} a_{k s_{t+1} s_{t+2}}}. \tag{3.38}$$

O desenvolvimento de um método alternativo de simulação para obtenção de realizações da distribuição $\pi(\mathbf{s}|\mathbf{y}, A, \mathbf{P})$ apresentado nos modelos anteriores e que se mostrou, através de estudos de simulação, mais eficiente que o método anteriormente discutido não foi possível para este modelo, pois o HMM$(2,1)$ apresenta estrutura de dependência mais complexa na cadeia de Markov oculta e S_t, para $t = 3, ..., T$, depende diretamente de S_{t-1} e S_{t-2} e não apenas de uma única variável.

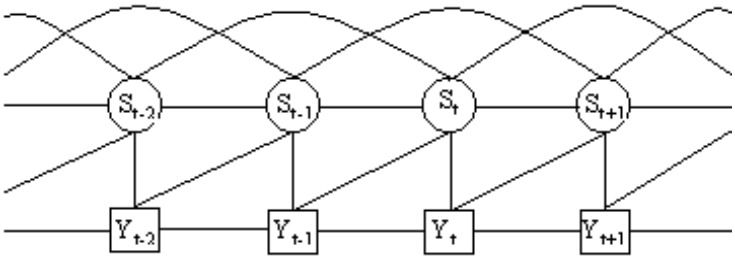


Figura 3.7: Grafo moral

Capítulo 4

Seleção de Modelos

Nas últimas décadas, os modelos HMMs, apresentados no capítulo anterior, têm sido aplicados com êxito em uma variedade de áreas de pesquisa, dentre as quais se destaca a Biologia Molecular, na identificação de segmentos homogêneos em seqüências de DNA (Boys *et al.*, 2004, 2002 e 2000). No entanto, geralmente, a ordem de dependência entre os estados observáveis ou ocultos é desconhecida, ou seja, a estrutura de dependência entre as variáveis, que especifica o modelo, não é conhecida e, assim, inferências sobre ela através dos dados é necessária.

A metodologia proposta por Friedman e Koller (2003) para estimação da estrutura de uma rede probabilística descrita na Seção 2.2 é uma alternativa para identificação da estrutura quando os dados são completos e da ordem de dependência entre as variáveis. Contudo, o problema de estimar a ordem de dependência da cadeia de Markov também pode ser visto como um problema de seleção de modelos, através do qual escolhemos o modelo mais adequado aos dados observados dentro do conjunto de modelos mais prováveis com diferentes ordens de dependência.

Neste capítulo, apresentamos dois métodos para seleção de modelos: o fator de Bayes, descrito em Kass e Raftery (1995) e o DIC (do inglês *deviance information criterion*), proposto em Spigelhalter *et al.* (2002).

4.1 Fator de Bayes

Seja $\mathbf{Y}$ um conjunto de observações produzido sob um de dois modelos M_0 ou M_1 de acordo com a densidade de probabilidades $\Pr(\mathbf{Y}|M_0)$ ou $\Pr(\mathbf{Y}|M_1)$. Dadas as probabilidades *a priori* $\Pr(M_0)$ e $\Pr(M_1) = 1 - \Pr(M_0)$ e considerando a abordagem bayesiana, temos as probabilidades *a posteriori* $\Pr(M_0|\mathbf{Y})$ e $\Pr(M_1|\mathbf{Y}) = 1 - \Pr(M_0|\mathbf{Y})$.

Através do teorema de Bayes, temos

$$\Pr(M_k|\mathbf{Y}) = \frac{\Pr(\mathbf{Y}|M_k)\Pr(M_k)}{\Pr(\mathbf{Y}|M_0)\Pr(M_0) + \Pr(\mathbf{Y}|M_1)\Pr(M_1)}, \tag{4.1}$$

para $k = 0, 1$. Assim,

$$\frac{\Pr(M_0|\mathbf{Y})}{\Pr(M_1|\mathbf{Y})} = \frac{\Pr(\mathbf{Y}|M_0)\Pr(M_0)}{\Pr(\mathbf{Y}|M_1)\Pr(M_1)}$$

e o fator de Bayes B_{01} pode ser definido como

$$B_{01} = \frac{\Pr(\mathbf{Y}|M_0)}{\Pr(\mathbf{Y}|M_1)}. \tag{4.2}$$

As densidades $\Pr(\mathbf{Y}|M_k)$, para $k = 0, 1$, podem ser obtidas pela integração sobre o espaço paramétrico, temos

$$\Pr(\mathbf{Y}|M_k) = \int \Pr(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)\pi(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k, \tag{4.3}$$

onde $\boldsymbol{\theta}_k$ é o vetor paramétrico sob M_k, $\pi(\boldsymbol{\theta}_k|M_k)$ é a sua densidade *a priori* e $\Pr(\mathbf{Y}|\boldsymbol{\theta}_k, M_k)$ é a densidade de probabilidades de $\mathbf{Y}$ dado o valor $\boldsymbol{\theta}_k$ ou a função de verossimilhança de $\boldsymbol{\theta}_k$.

Quando vários modelos são envolvidos na comparação, nós podemos escrever B_{jk} como o fator de Bayes para M_j contra M_k.

Seja M_0 o HMM com dependência de Markov de primeira ordem entre os estados ocultos e observáveis descrito na Seção 3.3 e M_1 o HMM com dependência de Markov de segunda ordem na seqüência de estados observáveis descrito na Seção 3.4, o fator de Bayes B_{01} pode ser definido como

$$B_{01} = \frac{\int\limits_{A,\mathbf{P}} \frac{1}{NM} \left(\prod_{k=1}^{N} \prod_{l=1}^{N} a_{kl}^{m_{kl1}+\beta_{kl}-1} \right) \left(\prod_{k=1}^{N} \prod_{i=1}^{M} \prod_{j=1}^{M} \left(p_{ij}^{(k)} \right)^{n_{ij}^{(k)}+\alpha_{ij}^{(k)}-1} \right) dA d\mathbf{P}}{\int\limits_{A,\mathbf{P}} \frac{1}{(NM)^2} \left(\prod_{k=1}^{N} \prod_{l=1}^{N} a_{kl}^{m_{kl2}+\beta_{kl}-1} \right) \left(\prod_{k=1}^{N} \prod_{h=1}^{M} \prod_{i=1}^{M} \prod_{j=1}^{M} \left(p_{hij}^{(k)} \right)^{n_{hij}^{(k)}+\alpha_{hij}^{(k)}-1} \right) dA d\mathbf{P}}$$

$$= \frac{\frac{1}{NM} \left(\prod_{k=1}^{N} \int\limits_{\mathbf{a}_k} \prod_{l=1}^{N} a_{kl}^{m_{kl1}+\beta_{kl}-1} d\mathbf{a}_k \right) \left(\prod_{k=1}^{N} \prod_{i=1}^{M} \int\limits_{\mathbf{p}_i^{(k)}} \prod_{j=1}^{M} \left(p_{ij}^{(k)} \right)^{n_{ij}^{(k)}+\alpha_{ij}^{(k)}-1} d\mathbf{p}_i^{(k)} \right)}{\frac{1}{(NM)^2} \left(\prod_{k=1}^{N} \int\limits_{\mathbf{a}_k} \prod_{l=1}^{N} a_{kl}^{m_{kl2}+\beta_{kl}-1} d\mathbf{a}_k \right) \left(\prod_{k=1}^{N} \prod_{h=1}^{M} \prod_{i=1}^{M} \int\limits_{\mathbf{p}_{hi}^{(k)}} \prod_{j=1}^{M} \left(p_{hij}^{(k)} \right)^{n_{hij}^{(k)}+\alpha_{hij}^{(k)}-1} d\mathbf{p}_{hi}^{(k)} \right)}$$

$$= NM \frac{\left(\prod_{k=1}^{N} \frac{\prod_{l=1}^{N} \Gamma(m_{kl1}+\beta_{kl})}{\Gamma\left(\sum_{l=1}^{N} (m_{kl1}+\beta_{kl}) \right)} \right) \left(\prod_{k=1}^{N} \prod_{i=1}^{M} \frac{\prod_{j=1}^{M} \Gamma\left(n_{ij}^{(k)}+\alpha_{ij}^{(k)} \right)}{\Gamma\left(\sum_{j=1}^{M} \left(n_{ij}^{(k)}+\alpha_{ij}^{(k)} \right) \right)} \right)}{\left(\prod_{k=1}^{N} \frac{\prod_{l=1}^{N} \Gamma(m_{kl2}+\beta_{kl})}{\Gamma\left(\sum_{l=1}^{N} (m_{kl2}+\beta_{kl}) \right)} \right) \left(\prod_{k=1}^{N} \prod_{h=1}^{M} \prod_{i=1}^{M} \frac{\prod_{j=1}^{M} \Gamma\left(n_{hij}^{(k)}+\alpha_{hij}^{(k)} \right)}{\Gamma\left(\sum_{j=1}^{M} \left(n_{hij}^{(k)}+\alpha_{hij}^{(k)} \right) \right)} \right)}, \tag{4.4}$$

onde

$$m_{kl1} = \sum_{t=1}^{T-1} I\left(s_t = k, s_{t+1} = l\right),$$

$$n_{ij}^{(k)} = \sum_{t=2}^{T} I\left(y_{t-1} = i, y_t = j, s_t = k\right),$$

$$m_{kl2} = \sum_{t=2}^{T-1} I\left(s_t = k, s_{t+1} = l\right) \text{ e}$$

$$n_{hij}^{(k)} = \sum_{t=3}^{T} I\left(y_{t-2} = h, y_{t-1} = i, y_t = j, s_t = k\right).$$

O fator de Bayes é um resumo da evidência fornecida pelos dados em favor de um modelo estatístico e em detrimento de outro. Considerando duas vezes o logaritmo natural do fator de Bayes, que está na mesma escala da deviance e do teste da razão de

verossimilhança, Kass e Raftery (1995) propõem a interpretação para o fator apresentada na Tabela 4.1.

Tabela 4.1. Interpretação do fator de Bayes

$2\log(B_{10})$	(B_{10})	evidência contra M_0
0 a 2	1 a 3	fraca
2 a 6	3 a 20	positiva
6 a 10	20 a 150	forte
>10	>150	muito forte

4.2 *DIC*

O critério proposto por Spigelhalter *et al.*(2002) para comparação e seleção de modelos combina medidas de ajuste (deviance estatística) e complexidade (número de parâmetros a serem estimados) do modelo com a finalidade de identificar o modelo que melhor explica os dados observados.

Spiegelhalter *et al.*(2002) propõem o DIC, *deviance information criterion*, definido como uma estimativa clássica do ajuste mais o dobro do número de parâmetros efetivos, dado por

$$DIC = D(\bar{\boldsymbol{\theta}}) + 2p_D, \tag{4.5}$$

onde $D(\boldsymbol{\theta}) = -2\log(\Pr(\mathbf{Y}|\boldsymbol{\theta})) - 2\log(h(\mathbf{y}))$, $\bar{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{Y})$ é a média *a posteriori* dos parâmetros, $h(\mathbf{y})$ é uma função dos dados e

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \tag{4.6}$$

é o número efetivo de parâmetros no modelo, calculado como a deviance média menos a deviance da média *a posteriori* dos parâmetros.

Usando (4.6), o DIC pode ser reescrito como

$$DIC = 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}). \tag{4.7}$$

O DIC pode ser calculado durante o processo de estimação através dos métodos

MCMC, monitorando os parâmetros estimados para $\boldsymbol{\theta}$ e, a cada iteração do processo, estimando $D(\boldsymbol{\theta})$. No final das iterações, calculamos a média amostral dos valores estimados de $D(\boldsymbol{\theta})$ e subtraímos a estimativa da deviance calculada usando as médias amostrais dos valores simulados de $\boldsymbol{\theta}$.

O modelo que melhor explica os dados observados é o modelo que apresenta o menor valor de DIC.

Capítulo 5

Simulações

Neste capítulo, apresentamos a aplicação dos HMMs com dependência de Markov de primeira ordem, descrito na Seção 3.3, e segunda ordem, discutido na Seção 3.4, em dois conjuntos de dados simulados e, para identificarmos o modelo mais adequado aos dados entre os dois estimados, utilizamos os métodos de seleção de modelos apresentados no Capítulo 4. Nosso principal objetivo é analisar a performance dos estimadores e dos métodos de seleção de modelos.

O primeiro conjunto de dados considerado foi gerado através de um modelo com dependência de primeira ordem entre os estados observáveis, HMM$(1, 1)$, enquanto que o segundo conjunto foi criado através de um HMM com dependência de segunda ordem entre os estados observáveis, HMM$(1, 2)$. Em ambas as situações, a seqüência de estados ocultos foi fixa e especificada pelos autores.

5.1 Dados com dependência de Markov de primeira ordem

Considerando uma seqüência de tamanho $T = 3000$, simulamos a seqüência de estados observáveis a partir de um modelo Markoviano oculto com $N = 2$ estados ocultos, $M = 4$ estados observáveis (resultados), dependência de Markov de primeira ordem e a seguinte seqüência de estados ocultos

$$\mathbf{s} = \underbrace{\{1,...,1,}_{1000} \underbrace{2,...,2,}_{1000} \underbrace{1,...,1\}}_{1000},$$

cujos pontos de mudança entre os estados ocultos foram fixados nas posições 1001 e 2001. A seguir, apresentamos as matrizes de transição entre os estados observáveis usadas para gerar a seqüência observada:

$$P^{(1)} = \begin{pmatrix} 0.098 & 0.457 & 0.075 & 0.370 \\ 0.121 & 0.557 & 0.202 & 0.120 \\ 0.059 & 0.410 & 0.087 & 0.444 \\ 0.453 & 0.350 & 0.067 & 0.130 \end{pmatrix} \quad \text{e} \quad P^{(2)} = \begin{pmatrix} 0.430 & 0.089 & 0.387 & 0.094 \\ 0.351 & 0.122 & 0.091 & 0.436 \\ 0.399 & 0.019 & 0.415 & 0.167 \\ 0.280 & 0.180 & 0.327 & 0.213 \end{pmatrix}.$$

Na seqüência observada aplicamos as metodologias de estimação do HMM$(1,1)$ e HMM$(1,2)$. O conhecimento *a priori* sobre as probabilidades de transição entre os estados observáveis, em cada estado oculto, considerado neste estudo de simulação foi não informativo e, por isso, escolhemos os hiperparâmetros das distribuições Dirichlet envolvidas nos modelos como $\boldsymbol{\alpha}_i^{(k)} = (1,1,1,1)$ e $\boldsymbol{\alpha}_{hi}^{(k)} = (1,1,1,1)$, para $i = 1,...,4$, $h = 1,...,4$ e $k = 1,2$. Desta maneira, cada componente $p_{ij}^{(k)}$ou $p_{hij}^{(k)}$, definidas respectivamente em (3.9) e (3.21), para $j = 1,...,4$, tem média 0.25 e desvio padrão, aproximadamente, 0.19. Para a matriz de transição entre os estados ocultos assumimos distribuições Dirichlet informativas, com hiperparâmetros $\beta_{11} = \beta_{22} = 99$ e $\beta_{12} = \beta_{12} = 2$. Assim, $E(a_{11}) = E(a_{22}) = 0.98$ e $Var(a_{11}) = Var(a_{22}) = 0.0002$, para a_{kk} definido em (3.8) e (3.20) e $k = 1,2$.

5.1.1 Resultados

Para cada uma das duas metodologias foram simuladas 11000 iterações com as primeiras 1000 sendo descartadas como *burn-in*. Nossos resultados são baseados em uma amostra de tamanho 2000, composta pelos valores gerados a cada 5 iterações. O diagnóstico de convergência Gelman-Rubin (Gelman e Rubin, 1992) foi utilizado para verificação de convergência do algoritmo. Para isto, duas cadeias com diferentes seqüências de estados

ocultos iniciais foram simuladas. As seqüências iniciais foram:

$$\mathbf{s}_0 = \underbrace{\{1,...,1,}_{500}\ \underbrace{2,...,2,}_{500}\ \underbrace{1,...,1,}_{500}\ \underbrace{2,...,2,}_{500}\ \underbrace{1,...,1,}_{500}\ \underbrace{2,...,2\}}_{500} \text{ e}$$

$$\mathbf{s}_0 = \underbrace{\{1,...,1\}}_{3000}.$$

As Tabelas 5.1 e 5.2 contêm, respectivamente, as estimativas do diagnóstico de Gelman-Rubin para as probabilidades de transição envolvidas no HMM$(1,1)$ e HMM$(1,2)$. Devido ao grande número de probabilidades de transição entre os estados observáveis (32 no HMM$(1,1)$ e 128 no HMM$(1,2)$), selecionamos, aleatoriamente, algumas estimativas do diagnóstico para exibir nas tabelas. Estas mostram que não há evidência de falta de convergência, pois as estimativas do diagnóstico Gelman-Rubin são menores que 1.1.

Tabela 5.1. Estimativas do diagnóstico Gelman-Rubin (GR) do HMM$(1,1)$

	GR		GR		GR
a_{11}	1.03	$p_{14}^{(1)}$	1.01	$p_{23}^{(2)}$	1.07
a_{12}	1.03	$p_{14}^{(2)}$	1.01	$p_{32}^{(1)}$	1.00
a_{21}	0.999	$p_{21}^{(2)}$	1.01	$p_{34}^{(2)}$	1.00
a_{22}	0.999	$p_{22}^{(1)}$	0.995	$p_{42}^{(2)}$	1.10

Tabela 5.2. Estimativas do diagnóstico Gelman-Rubin (GR) do HMM$(1,2)$

	GR		GR		GR		GR		GR
a_{11}	1.04	$p_{141}^{(2)}$	1.09	$p_{241}^{(2)}$	0.992	$p_{343}^{(2)}$	0.995	$p_{432}^{(1)}$	0.99
a_{12}	1.04	$p_{142}^{(1)}$	1.06	$p_{242}^{(2)}$	1.01	$p_{344}^{(1)}$	0.998	$p_{433}^{(1)}$	1.01
a_{21}	1.04	$p_{212}^{(1)}$	1.00	$p_{323}^{(1)}$	0.994	$p_{344}^{(2)}$	1.00	$p_{434}^{(2)}$	0.994
a_{22}	1.04	$p_{221}^{(1)}$	1.01	$p_{324}^{(1)}$	1.00	$p_{414}^{(2)}$	1.03	$p_{441}^{(1)}$	1.00
$p_{111}^{(2)}$	0.996	$p_{223}^{(1)}$	1.08	$p_{341}^{(1)}$	1.05	$p_{421}^{(1)}$	1.02	$p_{442}^{(1)}$	1.03
$p_{114}^{(1)}$	1.02	$p_{234}^{(1)}$	1.01	$p_{342}^{(2)}$	1.00	$p_{423}^{(2)}$	1.03	$p_{443}^{(2)}$	1.04

Através da Tabela 5.3, que contém as estimativas do fator de Bayes (B_{01}) e do DIC para os dois modelos, observamos que ambos os métodos selecionaram o HMM$(1,1)$ como o modelo mais adequado ao conjunto de dados, pois $B_{01} > 1$, $2\log(B_{01}) > 150$

e o DIC do HMM$(1,1)$ é menor que o DIC do HMM$(1,2)$. Este resultado indica boa performance dos métodos de seleção de modelos em determinar a ordem de dependência na seqüência de estados observáveis, considerando que os dados foram gerados de um HMM$(1,1)$.

Tabela 5.3. Estimativas dos métodos de seleção de modelos

Fator de Bayes (B_{01})	$1.10 * 10^{47}$
$2\log(B_{01})$	216.634
DIC do HMM$(1,1)$	6983.53
DIC do HMM$(1,2)$	7033.64

A Tabela 5.4 contém a média *a posteriori* das probabilidades de transição do modelo HMM$(1,1)$, através da qual notamos que os valores estimados não são muito diferentes dos valores a partir dos quais os dados foram gerados. Os pontos de mudança dos estados ocultos na seqüência estimada foram 1001 e 1999, evidenciando um acerto de 99.93% na identificação da seqüência dos estados ocultos.

Tabela 5.4. Médias *a posteriori* para as matrizes de transição do HMM$(1,1)$

$$\widehat{A} = \begin{pmatrix} 0.999 & 0.001 \\ 0.003 & 0.997 \end{pmatrix}$$

$$\widehat{P^{(1)}} = \begin{pmatrix} 0.089 & 0.415 & 0.062 & 0.433 \\ 0.124 & 0.564 & 0.193 & 0.118 \\ 0.047 & 0.408 & 0.094 & 0.450 \\ 0.439 & 0.380 & 0.068 & 0.113 \end{pmatrix} \quad \widehat{P^{(2)}} = \begin{pmatrix} 0.447 & 0.088 & 0.396 & 0.069 \\ 0.344 & 0.199 & 0.099 & 0.357 \\ 0.360 & 0.021 & 0.411 & 0.206 \\ 0.300 & 0.136 & 0.317 & 0.243 \end{pmatrix}$$

5.2 Dados com dependência de Markov de segunda ordem

Assim como na seção anterior, consideramos uma seqüência de tamanho $T = 3000$ e simulamos a seqüência de estados observáveis a partir de um modelo Markoviano oculto com $N = 2$ estados ocultos, $M = 4$ estados observáveis (resultados), dependência de Markov de segunda ordem e a seguinte seqüência de estados ocultos

$$\mathbf{s} = \underbrace{\{1,...,1,}_{1000} \underbrace{2,...,2,}_{1000} \underbrace{1,...,1\}}_{1000},$$

cujos pontos de mudança entre os estados ocultos foram fixados nas posições 1001 e 2001. A seguir, apresentamos as matrizes de transição entre os estados observáveis usadas para gerar a seqüência observada:

$$P^{(1)}_{1..} = \begin{pmatrix} 0.098 & 0.457 & 0.075 & 0.370 \\ 0.121 & 0.557 & 0.202 & 0.120 \\ 0.059 & 0.410 & 0.087 & 0.444 \\ 0.453 & 0.350 & 0.067 & 0.130 \end{pmatrix}, \quad P^{(1)}_{2..} = \begin{pmatrix} 0.430 & 0.089 & 0.387 & 0.094 \\ 0.351 & 0.122 & 0.091 & 0.436 \\ 0.399 & 0.019 & 0.415 & 0.167 \\ 0.280 & 0.180 & 0.327 & 0.213 \end{pmatrix},$$

$$P^{(1)}_{3..} = \begin{pmatrix} 0.30 & 0.40 & 0.10 & 0.20 \\ 0.10 & 0.20 & 0.30 & 0.40 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.35 & 0.15 & 0.20 & 0.30 \end{pmatrix}, \quad P^{(1)}_{4..} = \begin{pmatrix} 0.21 & 0.29 & 0.17 & 0.33 \\ 0.10 & 0.35 & 0.15 & 0.40 \\ 0.30 & 0.15 & 0.35 & 0.20 \\ 0.27 & 0.31 & 0.19 & 0.23 \end{pmatrix},$$

$$P^{(2)}_{1..} = \begin{pmatrix} 0.235 & 0.289 & 0.174 & 0.302 \\ 0.042 & 0.447 & 0.017 & 0.494 \\ 0.171 & 0.258 & 0.242 & 0.329 \\ 0.036 & 0.390 & 0.056 & 0.518 \end{pmatrix}, \quad P^{(2)}_{2..} = \begin{pmatrix} 0.353 & 0.159 & 0.232 & 0.256 \\ 0.341 & 0.208 & 0.042 & 0.409 \\ 0.325 & 0.198 & 0.217 & 0.260 \\ 0.233 & 0.166 & 0.208 & 0.393 \end{pmatrix},$$

$$P^{(2)}_{3..} = \begin{pmatrix} 0.318 & 0.251 & 0.121 & 0.310 \\ 0.330 & 0.198 & 0.368 & 0.104 \\ 0.106 & 0.107 & 0.726 & 0.061 \\ 0.092 & 0.586 & 0.040 & 0.282 \end{pmatrix} \text{ e } P^{(2)}_{4..} = \begin{pmatrix} 0.15 & 0.45 & 0.25 & 0.15 \\ 0.40 & 0.10 & 0.30 & 0.20 \\ 0.50 & 0.20 & 0.05 & 0.25 \\ 0.20 & 0.25 & 0.15 & 0.40 \end{pmatrix}.$$

Na seqüência observada aplicamos as metodologias de estimação do HMM$(1,1)$

e HMM$(1,2)$. O conhecimento *a priori* sobre as probabilidades de transição entre os estados observáveis considerado neste estudo de simulação foi não informativo e, por isso, escolhemos os hiperparâmetros das distribuições Dirichlet envolvidas nos modelos como $\boldsymbol{\alpha}_i^{(k)} = (1,1,1,1)$ e $\boldsymbol{\alpha}_{hi}^{(k)} = (1,1,1,1)$, para $i = 1,...,4$, $h = 1,...,4$ e $k = 1,2$. Desta maneira, cada componente $p_{ij}^{(k)}$ou $p_{hij}^{(k)}$, para $j = 1,...,4$, tem média 0.25 e desvio padrão, aproximadamente, 0.19. Para a matriz de transição entre os estados ocultos assumimos distribuições Dirichlet informativas, com hiperparâmetros $\beta_{11} = \beta_{22} = 99$ e $\beta_{12} = \beta_{12} = 2$. Assim, $E(a_{11}) = E(a_{22}) = 0.98$ e $Var(a_{11}) = Var(a_{22}) = 0.0002$.

5.2.1 Resultados

Para cada uma das duas metodologias foram simuladas 11000 iterações com as primeiras 1000 sendo descartadas como *burn-in*. Nossos resultados são baseados em uma amostra de tamanho 2000, composta pelos valores gerados a cada 5 iterações. O diagnóstico de convergência Gelman-Rubin (Gelman e Rubin, 1992) foi utilizado para verificação de convergência do algoritmo. Para isto, duas cadeias com diferentes seqüências de estados ocultos iniciais foram simuladas. As seqüências iniciais foram

$$\mathbf{s}_0 = \underbrace{\{1,...,1,}_{500}\ \underbrace{2,...,2,}_{500}\ \underbrace{1,...,1,}_{500}\ \underbrace{2,...,2,}_{500}\ \underbrace{1,...,1,}_{500}\ \underbrace{2,...,2\}}_{500} \text{ e}$$

$$\mathbf{s}_0 = \underbrace{\{1,...,1\}}_{3000}.$$

As Tabelas 5.5 e 5.6 contêm, respectivamente, as estimativas do diagnóstico de Gelman-Rubin para as probabilidades de transição envolvidas no HMM$(1,1)$ e HMM$(1,2)$. Devido ao grande número de probabilidades de transição entre os estados observáveis (32 no HMM$(1,1)$ e 128 no HMM$(1,2)$), selecionamos, aleatoriamente, algumas estimativas do diagnóstico para exibir nas tabelas.

Na Tabela 5.5 observamos que os valores do diagnóstico de Gelman-Rubin são maiores que 1.1, fato que indica a não convergência das estimativas dos parâmetros envolvidos no HMM$(1,1)$, ou seja, a quantidade de iterações simuladas, o *burn-in* e o salto na coleta da amostra talvez não tenham sido suficientes para garantir a convergência

do algoritmo. Por este motivo, não podemos utilizar as estimativas do HMM$(1,1)$ nos métodos de seleção de modelos para comparar e selecionar o HMM mais adequado aos dados.

Tabela 5.5. Estimativas do diagnóstico Gelman-Rubin (GR) do HMM$(1,1)$

	GR		GR		GR
a_{11}	3.28	$p_{13}^{(2)}$	4.09	$p_{32}^{(1)}$	3.18
a_{12}	3.28	$p_{14}^{(2)}$	3.43	$p_{33}^{(2)}$	4.91
a_{21}	4.37	$p_{21}^{(2)}$	5.06	$p_{34}^{(1)}$	6.60
a_{22}	4.37	$p_{23}^{(1)}$	2.21		

Em contrapartida, através da Tabela 5.6 observamos que não há evidência de falta de convergência nas estimativas do HMM$(1,2)$.

Tabela 5.6. Estimativas do diagnóstico Gelman-Rubin (GR) do HMM$(1,2)$

	GR		GR		GR		GR		GR
a_{11}	0.99	$p_{121}^{(2)}$	0.998	$p_{144}^{(1)}$	1.02	$p_{234}^{(2)}$	1.01	$p_{343}^{(2)}$	1.03
a_{12}	0.99	$p_{122}^{(2)}$	1.08	$p_{211}^{(2)}$	1.05	$p_{2422}^{(2)}$	0.997	$p_{411}^{(2)}$	0.992
a_{21}	0.99	$p_{114}^{(2)}$	1.10	$p_{212}^{(1)}$	1.08	$p_{311}^{(1)}$	1.05	$p_{412}^{(2)}$	1.01
a_{22}	0.993	$p_{133}^{(2)}$	1.04	$p_{213}^{(1)}$	0.996	$p_{311}^{(2)}$	1.00	$p_{413}^{(1)}$	0.99
$p_{111}^{(2)}$	1.03	$p_{141}^{(1)}$	1.02	$p_{222}^{(2)}$	1.02	$p_{314}^{(2)}$	0.999	$p_{433}^{(2)}$	1.00
$p_{114}^{(2)}$	0.999	$p_{142}^{(1)}$	1.04	$p_{233}^{(2)}$	1.01	$p_{322}^{(1)}$	0.993	$p_{443}^{(1)}$	0.996

A Tabela 5.7 contém a média *a posteriori* das probabilidades de transição do modelo HMM$(1,2)$, através da qual notamos que os valores estimados não são muito diferentes dos valores a partir dos quais os dados foram gerados. Os pontos de mudança dos estados ocultos na seqüência estimada foram 1007 e 2002, evidenciando um acerto de 99.77% na identificação da seqüência dos estados ocultos.

Tabela 5.7. Médias a posteriori para as matrizes de transição do HMM$(1, 2)$

$$\widehat{A} = \begin{pmatrix} 0.998 & 0.002 \\ 0.003 & 0.997 \end{pmatrix}$$

$$\widehat{P^{(1)}_{1..}} = \begin{pmatrix} 0.099 & 0.439 & 0.084 & 0.377 \\ 0.152 & 0.562 & 0.177 & 0.109 \\ 0.067 & 0.336 & 0.105 & 0.490 \\ 0.409 & 0.402 & 0.067 & 0.121 \end{pmatrix} \quad \widehat{P^{(1)}_{2..}} = \begin{pmatrix} 0.467 & 0.076 & 0.370 & 0.087 \\ 0.328 & 0.139 & 0.100 & 0.432 \\ 0.382 & 0.009 & 0.455 & 0.154 \\ 0.261 & 0.191 & 0.354 & 0.193 \end{pmatrix}$$

$$\widehat{P^{(1)}_{3..}} = \begin{pmatrix} 0.290 & 0.368 & 0.090 & 0.253 \\ 0.091 & 0.171 & 0.334 & 0.404 \\ 0.243 & 0.220 & 0.299 & 0.238 \\ 0.417 & 0.147 & 0.200 & 0.236 \end{pmatrix} \quad \widehat{P^{(1)}_{4..}} = \begin{pmatrix} 0.174 & 0.268 & 0.172 & 0.385 \\ 0.118 & 0.358 & 0.165 & 0.358 \\ 0.306 & 0.121 & 0.341 & 0.232 \\ 0.256 & 0.297 & 0.205 & 0.242 \end{pmatrix}$$

$$\widehat{P^{(2)}_{1..}} = \begin{pmatrix} 0.281 & 0.297 & 0.105 & 0.317 \\ 0.028 & 0.427 & 0.013 & 0.532 \\ 0.124 & 0.165 & 0.255 & 0.456 \\ 0.028 & 0.389 & 0.027 & 0.555 \end{pmatrix} \quad \widehat{P^{(2)}_{2..}} = \begin{pmatrix} 0.340 & 0.114 & 0.195 & 0.351 \\ 0.357 & 0.223 & 0.025 & 0.395 \\ 0.345 & 0.179 & 0.161 & 0.314 \\ 0.232 & 0.172 & 0.226 & 0.361 \end{pmatrix}$$

$$\widehat{P^{(2)}_{3..}} = \begin{pmatrix} 0.325 & 0.241 & 0.078 & 0.355 \\ 0.312 & 0.239 & 0.309 & 0.139 \\ 0.088 & 0.150 & 0.679 & 0.083 \\ 0.133 & 0.381 & 0.083 & 0.402 \end{pmatrix} \quad \widehat{P^{(2)}_{4..}} = \begin{pmatrix} 0.177 & 0.454 & 0.248 & 0.121 \\ 0.491 & 0.074 & 0.299 & 0.135 \\ 0.318 & 0.381 & 0.044 & 0.257 \\ 0.195 & 0.228 & 0.149 & 0.429 \end{pmatrix}$$

Capítulo 6

Aplicações

Um grande número de pesquisas dos últimos anos tem focalizado o entendimento e identificação da estrutura de dependência presente em uma seqüência de DNA. Neste capítulo, aplicamos os métodos descritos nas seções anteriores no intron 7 do gene $\alpha - fetoprotein$ dos chimpanzés e no genoma do *Bacteriophage lambda*.

6.1 Modelando uma seqüência de DNA como um HMM

A metodologia dos modelos Markovianos ocultos tem sido aplicada em várias áreas de pesquisa, principalmente na Biologia Molecular e Genética, para segmentação da seqüência de DNA ou localização de segmentos homogêneos dentro desta seqüência.

O ácido desoxirribonucléico (DNA) é um portador básico de informação genética e é encontrado em todas as células vivas. O DNA é constituído por uma seqüência de quatro bases nitrogenadas {Adenina (A), Citosina (C), Guanina (G) e Timina (T)}. Assim, uma seqüência de DNA $\mathbf{y} = y_1, y_2, ..., y_T$ pode ser vista como uma realização do processo aleatório observado $\mathbf{Y} = Y_1, Y_2, ..., Y_T$, onde $Y_t \in \{A, C, G, T\} \equiv \{1, 2, 3, 4\}$ para $t = 1, 2, ..., T$. Suponha que existam N tipos de segmentos homogêneos dentro da seqüência de DNA. O tipo de segmento não observado na posição t é representado por $S_t \in \{1, 2, ..., N\}$, para $t = 1, 2, ..., T$.

Geralmente, as análises realizadas através do HMM para modelar a heterogeneidade presente na composição da seqüência de DNA assumem que a seqüência observada é composta por observações condicionalmente independentes dada a seqüência dos estados

ocultos. No entanto, evidências empíricas sugerem que este modelo não é suficiente para capturar e modelar a estrutura de dependência rica e complexa do DNA. Nestas circunstâncias, os métodos descritos neste trabalho podem ser usados para estimar a ordem de dependência das bases.

Como a metodologia proposta por Friedman e Koller (2003) para estimação da estrutura de dependência entre um conjunto de variáveis aleatórias apresentou desempenho insatisfatório nas situações consideradas e os pesquisadores biólogos acreditam que os modelos com primeira, HMM$(1,1)$, ou segunda, HMM$(1,2)$, ordens de dependência são suficientes para modelar seqüências de DNA, o modelo mais adequado aos dados, entre os estudados, pode ser identificado através de métodos de seleção de modelos, tais como: DIC e fator de Bayes, descritos no Capítulo 4.

Ilustramos o uso dos métodos de seleção e dos métodos de estimação do HMM$(1,1)$ e HMM$(1,2)$ apresentados nas seções anteriores através da análise do intron 7 do gene $\alpha - fetoprotein$ dos chimpanzés (Boys *et al.*, 2000 e 2002) e do genoma do parasita *Bacteriophage lambda* (Braun e Müller, 1998; Churchill, 1989; da-Silva, 2003 e Boys e Henderson, 2004).

6.2 Informações *a priori* sobre as probabilidades de transição entre as bases em cada segmento

Geralmente, temos pouco conhecimento sobre as probabilidades de transição entre as bases em cada segmento e um dos nossos principais objetivos (identificar e localizar os possíveis segmentos homogêneos) se relaciona diretamente com os estados ocultos e não com os observáveis. Desta maneira, devemos usar distribuições *a priori*, no caso Dirichlet, pouco informativas para $\mathbf{P}$ e, então, escolhemos $\boldsymbol{\alpha}_i^{(k)} = (1,1,1,1)$ e $\boldsymbol{\alpha}_{hi}^{(k)} = (1,1,1,1)$, para $i = 1,...,4$, $h = 1,...,4$ e $k = 1,2,...,N$. Nestes casos, as componentes $p_{ij}^{(k)}$ e $p_{hij}^{(k)}$, definidas em (3.9) e (3.21), para $j = 1,...,4$, seguem distribuição *a priori* Beta$(1,3)$, média 0.25 e desvio padrão, aproximadamente, 0.19. A correlação entre os elementos da linha é próxima a -0.47.

6.3 Informações *a priori* sobre as probabilidades de transição entre os segmentos homogêneos

A especificação da distribuição *a priori* da matriz A de transição entre os estados ocultos é mais elaborada. Contudo, considerando o comprimento dos segmentos, temos mais conhecimento sobre o comportamento destas transições.

Em geral, não é realístico tentarmos descobrir pequenos segmentos dentro da seqüência de DNA, exceto quando procuramos por segmentos de uma dada matriz de transição de bases ou quando há muitos fragmentos pequenos de um tipo de segmento particular. Assim, assumimos que transições entre tipos de segmentos são raras e $E(a_{kk})$, para a_{kk} definido em (3.8) e (3.20) e $k = 1, 2, ..., N$, são próximas a 1. Assumimos também que os elementos a_{kl} $(k \neq l)$ fora da diagonal principal são permutáveis dentro da linha e que os parâmetros da distribuição *a priori* Dirichlet têm a forma $\boldsymbol{\beta}_k = \{d, d, ..., d, c, d, d, ..., d\}$, onde c é o k-ésimo elemento de $\boldsymbol{\beta}_k$, para $k = 1, 2, ..., N$, e o número de elementos em $\boldsymbol{\beta}_k$ é N.

Portanto, os elementos da diagonal a_{kk} têm distribuição *a priori* Beta$(c, (N-1)d)$, com média *a priori* igual $\frac{c}{c+(N-1)d}$ e desvio padrão

$$\frac{\sqrt{\frac{(N-1)cd}{(c+(N-1)d+1)}}}{c+(N-1)d}.$$

6.4 Intron 7 do gene $\alpha - fetoprotein$ dos chimpanzés

O intron 7 do gene $\alpha - fetoprotein$ dos chimpanzés, analisado por Boys *et al.* (2000) através do HMM$(1, 1)$, é uma seqüência com $T = 1968$ pares de bases e este gene é um importante fator do desenvolvimento embrionário nos mamíferos e também parece interferir no desenvolvimento de tumores.

Para comparação com os resultados em Boys *et al.* (2000, 2002), assumimos $N = 3$ estados ocultos.

Na seqüência observada, aplicamos as metodologias de estimação do HMM$(1, 1)$ e HMM$(1, 2)$. O conhecimento *a priori* sobre as probabilidades de transição entre os estados observáveis, em cada estado oculto, considerado foi não informativo e, por isso,

escolhemos os hiperparâmetros das distribuições Dirichlet envolvidas nos modelos como $\boldsymbol{\alpha}_i^{(k)} = (1,1,1,1)$ e $\boldsymbol{\alpha}_{hi}^{(k)} = (1,1,1,1)$, para $i = 1,...,4$, $h = 1,...,4$ e $k = 1,2$ e 3. Para a matriz de transição entre os estados ocultos assumimos distribuições Dirichlet informativas, com hiperparâmetros $\beta_{kl} = 97.02$ para $k = l$ e $\beta_{kl} = 0.49$ caso contrário. Assim, $E(a_{kk}) = 0.99$ e $Var(a_{kk}) = 0.0001$.

6.4.1 Resultados

Para cada uma das duas metodologias foram simuladas 11000 iterações com as primeiras 1000 sendo descartadas como *burn-in*. Nossos resultados são baseados em uma amostra de tamanho 2000, composta pelos valores gerados a cada 5 iterações. O diagnóstico de convergência Gelman-Rubin (Gelman e Rubin, 1992) foi utilizado para verificação de convergência do algoritmo. Para isto, duas cadeias com diferentes seqüências de estados ocultos iniciais foram simuladas. As seqüências iniciais foram

$$\mathbf{s}_0 = \underbrace{\{1,...,1,}_{656} \underbrace{2,...,2,}_{656} \underbrace{3,...,3}_{656} \text{ e}$$

$$\mathbf{s}_0 = \underbrace{\{1,...,1\}}_{1968}.$$

As Tabelas 6.1 e 6.2 contêm, respectivamente, as estimativas do diagnóstico de Gelman-Rubin para as probabilidades de transição envolvidas no HMM$(1,1)$ e HMM$(1,2)$. Devido ao grande número de probabilidades de transição entre os estados observáveis (48 no HMM$(1,1)$ e 192 no HMM$(1,2)$), selecionamos, aleatoriamente, algumas estimativas do diagnóstico para exibir nas tabelas. Estas mostram que não há evidência de falta de convergência, pois as estimativas do diagnóstico Gelman-Rubin são menores que 1.1.

Tabela 6.1. Estimativas do diagnóstico Gelman-Rubin (GR) do HMM$(1,1)$

	GR		GR		GR
a_{11}	0.998	$p_{13}^{(1)}$	1.01	$p_{23}^{(1)}$	1.03
a_{13}	1.00	$p_{14}^{(2)}$	1.02	$p_{31}^{(3)}$	1.02
a_{22}	1.02	$p_{21}^{(3)}$	1.00	$p_{32}^{(1)}$	1.01
a_{32}	1.03	$p_{13}^{(3)}$	0.99	$p_{33}^{(2)}$	0.995

Tabela 6.2. Estimativas do diagnóstico Gelman-Rubin (GR) do HMM$(1,2)$

	GR		GR		GR		GR		GR
a_{11}	1.05	$p^{(3)}_{113}$	1.01	$p^{(2)}_{141}$	1.03	$p^{(1)}_{223}$	1.00	$p^{(3)}_{331}$	0.996
a_{12}	1.02	$p^{(2)}_{121}$	0.993	$p^{(2)}_{143}$	1.00	$p^{(3)}_{231}$	1.00	$p^{(2)}_{312}$	1.04
a_{22}	1.07	$p^{(2)}_{122}$	1.07	$p^{(3)}_{144}$	1.00	$p^{(3)}_{233}$	0.997	$p^{(1)}_{314}$	1.02
a_{33}	1.00	$p^{(2)}_{131}$	1.04	$p^{(2)}_{212}$	1.00	$p^{(2)}_{234}$	0.995	$p^{(2)}_{321}$	1.02
$p^{(1)}_{111}$	1.03	$p^{(3)}_{133}$	1.07	$p^{(2)}_{213}$	0.992	$p^{(2)}_{242}$	0.992	$p^{(3)}_{331}$	1.02
$p^{(2)}_{112}$	1.02	$p^{(2)}_{134}$	1.01	$p^{(2)}_{214}$	0.994	$p^{(1)}_{244}$	1.01	$p^{(2)}_{332}$	0.993

Através da Tabela 6.3, que contém as estimativas do fator de Bayes (B_{01}) e do DIC para os dois modelos, observamos que ambos os métodos selecionaram o HMM$(1,1)$ como o modelo mais adequado ao conjunto de dados, pois $B_{01} > 1$, $2\log(B_{01}) > 150$ e o DIC do HMM$(1,1)$, apesar de ser muito próximo, é menor que o DIC do HMM$(1,2)$. Este resultado concorda com a análise feita por Boys e Henderson (2002), que estimam a ordem de dependência no processo observado através de outra metodologia.

Tabela 6.3. Estimativas dos métodos de seleção de modelos

Fator de Bayes (B_{01})	$5.92 * 10^{42}$
$2\log(B_{01})$	196.97
DIC do HMM$(1,1)$	5196.96
DIC do HMM$(1,2)$	5206.41

Sumários *a posteriori* para as probabilidades de transição envolvidas no HMM$(1,1)$ são apresentados na Tabela 6.4, através da qual observamos que a matriz de transição média $P^{(2)}$ contém as probabilidades estimadas mais diferentes, com baixas probabilidades de observação das bases "A"e "G"e probabilidades de emissão de "C"e "T"mais elevadas.

Um gráfico útil para representar o processo dos estados ocultos pode ser construído com as realizações de $\mathbf{S}$ produzidas durante o processo de estimação. A probabilidade de ocorrência de cada estado oculto (tipo de segmento) em cada posição da seqüência, $\Pr(S_t = k|\mathbf{y})$, pode ser estimada pela proporção de iterações nas quais cada tipo de segmento ocorre em cada posição, ou seja,

$$\widehat{\Pr}(S_t = k|\mathbf{y}) = \frac{1}{2000}\sum_{i=1}^{2000} I\left(s_t^{(i)} = k\right),$$

para $t = 1, 2, ..., T$, $k = 1, 2, ..., N$ e onde 2000 é o número de amostras consideradas na estimação dos parâmetros e $I(A)$ é a função indicadora que assume o valor 1 se A é verdade e 0 caso contrário.

Tabela 6.4. Sumários *a posteriori* para as matrizes de transição do HMM$(1, 1)$

	Média	Desvio Padrão
A	$\begin{pmatrix} 0.997 & 0.001 & 0.002 \\ 0.008 & 0.989 & 0.003 \\ 0.011 & 0.004 & 0.985 \end{pmatrix}$	$\begin{pmatrix} 0.002 & 0.001 & 0.002 \\ 0.007 & 0.008 & 0.004 \\ 0.009 & 0.005 & 0.009 \end{pmatrix}$
$P^{(1)}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.35 & 0.16 & 0.23 & 0.26 \\ 0.34 & 0.21 & 0.04 & 0.41 \\ 0.32 & 0.20 & 0.21 & 0.26 \\ 0.23 & 0.17 & 0.20 & 0.39 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.02 & 0.02 & 0.02 & 0.02 \\ 0.03 & 0.02 & 0.01 & 0.03 \\ 0.03 & 0.02 & 0.03 & 0.03 \\ 0.02 & 0.02 & 0.02 & 0.02 \end{pmatrix} \end{array}$
$P^{(2)}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.21 & 0.31 & 0.15 & 0.32 \\ 0.04 & 0.45 & 0.01 & 0.50 \\ 0.13 & 0.26 & 0.24 & 0.37 \\ 0.03 & 0.39 & 0.05 & 0.53 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.17 & 0.18 & 0.13 & 0.18 \\ 0.03 & 0.06 & 0.02 & 0.06 \\ 0.12 & 0.15 & 0.14 & 0.17 \\ 0.02 & 0.06 & 0.03 & 0.06 \end{pmatrix} \end{array}$
$P^{(3)}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.35 & 0.11 & 0.49 & 0.05 \\ 0.34 & 0.23 & 0.13 & 0.30 \\ 0.33 & 0.19 & 0.35 & 0.12 \\ 0.18 & 0.12 & 0.61 & 0.09 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.12 & 0.05 & 0.13 & 0.03 \\ 0.11 & 0.09 & 0.07 & 0.10 \\ 0.06 & 0.06 & 0.07 & 0.05 \\ 0.11 & 0.08 & 0.12 & 0.07 \end{pmatrix} \end{array}$

A Figura 6.1 mostra a estimativa da probabilidade de cada estado oculto no decorrer da seqüência. Através dela, observamos que a primeira metade da seqüência é constituída, basicamente, por um único segmento, que se repete na região central da segunda metade da seqüência.

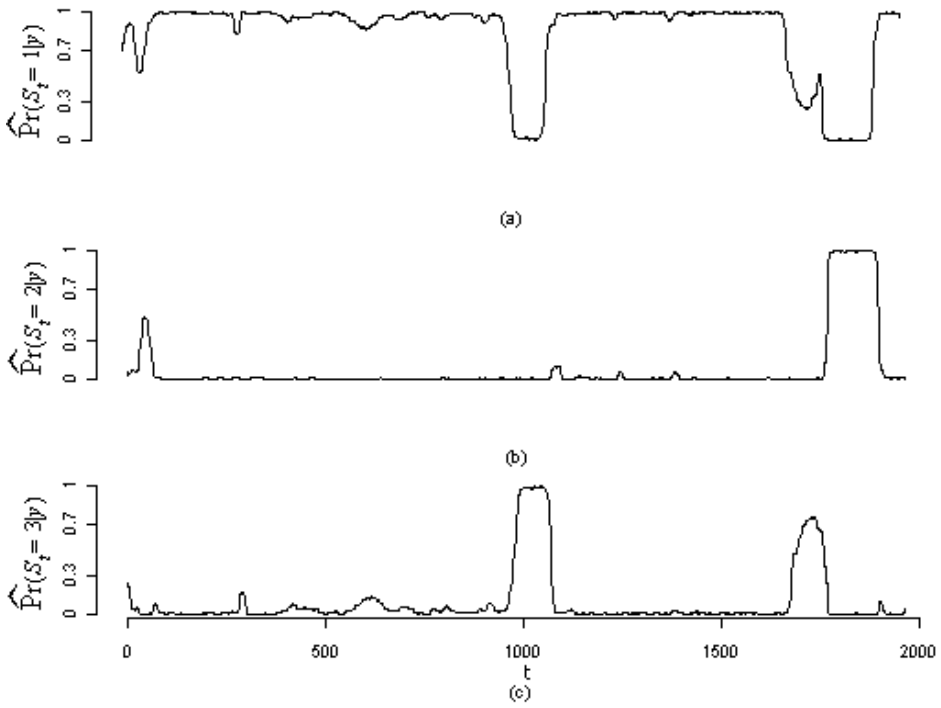


Figura 6.1: (a)-(c) Gráficos da probabilidade estimada *a posteriori* dos tipos de segmento

6.5 Genoma do *Bacteriophage lambda*

O *Bacteriophage lambda* é um parasita da bactéria intestinal *Escherichia coli*. O genoma deste vírus é constituído por $T = 48502$ pares de bases nitrogenadas no comprimento e sua seqüência genômica completa pode ser obtida na página http://www.ncbi.nlm.nih.gov/ do Centro Nacional de Informação Biotecnológica (NCBI).

Usando análise química, Skalka *et al.* (1968) concluíram que este genoma é composto por seis segmentos com diferentes proporções de $C + G$. Churchill (1989), por sua vez, comparou modelos com observações independentes e e com dependência de primeira ordem e determinou, através do critério de informação bayesiano (BIC), que o modelo com dependência de primeira ordem e $N = 3$ estados ocultos fornece melhor descrição dos dados. da-Silva (2003) apresenta os mesmos resultados de Churchill (1989).

Boys e Henderson (2004) propõem uma análise bayesiana da seqüência de DNA, assumindo um HMM no qual o processo observado envolve uma cadeia de Markov com número de estados ocultos e ordem de dependência, a princípio, desconhecidos e identificam o HMM$(1, 2)$ com $N = 6$ estados ocultos como o modelo mais adequado à seqüência

genômica do *Bacteriophage lambda*.

Baseados nos estudos anteriores, estimamos e comparamos, para esta seqüência, seis HMMs diferentes: HMM$(1,1)$ com $N = 2$ (M_1); HMM$(1,1)$ com $N = 3$ (M_2); HMM$(1,1)$ com $N = 4$ (M_3); HMM$(1,2)$ com $N = 2$ (M_4); HMM$(1,2)$ com $N = 3$ (M_5) e HMM$(1,2)$ com $N = 4$ (M_6).

O conhecimento *a priori* sobre as probabilidades de transição entre os estados observáveis, em cada estado oculto, considerado foi não informativo e, por isso, escolhemos os hiperparâmetros das distribuições Dirichlet envolvidas nos modelos como $\boldsymbol{\alpha}_i^{(k)} = (1,1,1,1)$ e $\boldsymbol{\alpha}_{hi}^{(k)} = (1,1,1,1)$, para $i = 1,...,4$, $h = 1,...,4$ e k variando de acordo com o número de estados ocultos em cada modelo considerado. Para a matriz de transição entre os estados ocultos assumimos distribuições Dirichlet informativas, com hiperparâmetros calculados para manter $E(a_{kk}) = 0.99$ e $Var(a_{kk}) = 0.0001$ em todos os modelos.

6.5.1 Resultados

Para cada um dos modelos foram simuladas 11000 iterações com as primeiras 1000 sendo descartadas como *burn-in*. Nossos resultados são baseados em uma amostra de tamanho 2000, composta pelos valores gerados a cada 5 iterações. A convergência do algoritmo MCMC foi monitorada através do uso de gráficos e diagnósticos de Gelman-Rubin.

A Tabela 6.5 contém as estimativas do DIC de cada modelo ajustado e a Tabela 6.6 apresenta as estimativas do fator de Bayes para cada par de modelos, onde B_{ij} representa o fator de Bayes de comparação entre os modelos M_i e M_j, para $i = 1,2,...,5$ e $j = i+1, i+2, ..., 6$.

Tabela 6.5. Estimativas do DIC

Modelo	DIC
M_1	132568.9
M_2	132699.7
M_3	132942.4
M_4	130936.6
M_5	131145.8
M_6	131017.4

Tabela 6.6. Estimativas do fator de Bayes

	B_{ij}	$2\log(B_{ij})$
B_{12}	$2.2 * 10^{11}$	52.2
B_{13}	$4.95 * 10^{59}$	274.9
B_{14}	$1.35 * 10^{-251}$	-1155.3
B_{15}	$3.72 * 10^{-242}$	-1111.8
B_{16}	$5.79 * 10^{-184}$	-843.8
B_{23}	$2.2 * 10^{48}$	222.6
B_{24}	$6.04 * 10^{-263}$	-1207.6
B_{25}	$1.66 * 10^{-253}$	-1164.1
B_{26}	$2.59 * 10^{-195}$	-448.1
B_{34}	$2.7 * 10^{-311}$	-1430.2
B_{35}	$7.52 * 10^{-302}$	-1386.7
B_{36}	$1.17 * 10^{-243}$	-1118.7
B_{45}	$3.67 * 10^{9}$	44.1
B_{46}	$4.28 * 10^{67}$	311.5
B_{56}	$1.16 * 10^{58}$	267.4

Através da Tabela 6.5, observamos que, pelo método de seleção DIC, o modelo M_4, HMM$(1, 2)$ com $N = 2$ segmentos homogêneos, foi selecionado como o modelo mais adequado ao genoma do parasita *Bacteriophage lambda*, pois a estimativa do DIC do M_4 é menor do que a estimativa do DIC dos demais modelos. Na Tabela 6.6, notamos o mesmo resultado observado na tabela anterior, ou seja, o modelo M_4, entre os modelos testados, também é apontado pelo fator de Bayes como o modelo mais adequado aos dados, pois as estimativas de todos os fatores de Bayes que comparam modelos de primeira com modelos de segunda ordem (B_{14}, B_{15}, B_{15}, B_{24}, B_{25}, B_{26}, B_{34}, B_{35} e B_{36}) evidenciam a favor do modelo de segunda ordem e, entre estes últimos, o modelo com $N = 2$ segmentos homogêneos é o mais adequado.

A Tabela 6.7 contém sumários *a posteriori* para as probabilidades de transição envolvidas no M_4, através da qual notamos que a principal diferença entre as estimativas médias para $P^{(1)}$ e $P^{(2)}$ são os valores das probabilidades de transição para a base nitrogenada "T", que são maiores no segundo segmento homogêneo. Vale destacar, tam-

bém, as probabilidades estimadas de observação da base "G"quando a base observada no penúltimo instante é a base "C"e o estado oculto é o primeiro segmento, cujos valores são altos comparados com as demais probabilidades estimadas.

Tabela 6.7. Sumários *a posteriori* para as matrizes de transição do M_4

	Média	Desvio Padrão
A	$\begin{pmatrix} 0.9997 & 0.0003 \\ 0.0004 & 0.9996 \end{pmatrix}$	$\begin{pmatrix} 0.0001 & 0.0001 \\ 0.0002 & 0.0002 \end{pmatrix}$
$P^{(1)}_{A..}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.35 & 0.24 & 0.23 & 0.17 \\ 0.24 & 0.27 & 0.33 & 0.16 \\ 0.25 & 0.31 & 0.26 & 0.18 \\ 0.16 & 0.25 & 0.38 & 0.22 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.010 & 0.010 & 0.009 & 0.009 \\ 0.010 & 0.011 & 0.011 & 0.009 \\ 0.010 & 0.011 & 0.010 & 0.009 \\ 0.009 & 0.011 & 0.012 & 0.010 \end{pmatrix} \end{array}$
$P^{(1)}_{C..}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.17 & 0.20 & 0.43 & 0.20 \\ 0.24 & 0.15 & 0.44 & 0.16 \\ 0.18 & 0.26 & 0.35 & 0.22 \\ 0.08 & 0.16 & 0.59 & 0.17 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.009 & 0.009 & 0.012 & 0.009 \\ 0.011 & 0.009 & 0.013 & 0.009 \\ 0.008 & 0.009 & 0.010 & 0.008 \\ 0.008 & 0.010 & 0.014 & 0.010 \end{pmatrix} \end{array}$
$P^{(1)}_{G..}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.32 & 0.23 & 0.19 & 0.26 \\ 0.28 & 0.23 & 0.28 & 0.21 \\ 0.27 & 0.31 & 0.21 & 0.22 \\ 0.18 & 0.22 & 0.39 & 0.22 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.010 & 0.009 & 0.008 & 0.009 \\ 0.009 & 0.008 & 0.009 & 0.008 \\ 0.009 & 0.009 & 0.008 & 0.008 \\ 0.009 & 0.010 & 0.012 & 0.010 \end{pmatrix} \end{array}$
$P^{(1)}_{T..}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.28 & 0.31 & 0.04 & 0.37 \\ 0.32 & 0.25 & 0.24 & 0.18 \\ 0.30 & 0.28 & 0.28 & 0.15 \\ 0.20 & 0.28 & 0.24 & 0.28 \end{pmatrix} \end{array}$	$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0.015 & 0.015 & 0.007 & 0.016 \\ 0.013 & 0.012 & 0.010 & 0.011 \\ 0.010 & 0.009 & 0.009 & 0.007 \\ 0.011 & 0.013 & 0.012 & 0.013 \end{pmatrix} \end{array}$

Continuação da Tabela 6.7

	Média					Desvio Padrão				
		A	C	G	T		A	C	G	T
$P^{(2)}_{A..}$	A	0.33	0.20	0.17	0.30	A	0.012	0.010	0.009	0.012
	C	0.30	0.25	0.19	0.26	C	0.016	0.015	0.014	0.015
	G	0.25	0.25	0.21	0.29	G	0.014	0.015	0.013	0.015
	T	0.24	0.21	0.23	0.32	T	0.011	0.010	0.010	0.012
		A	C	G	T		A	C	G	T
$P^{(2)}_{C..}$	A	0.29	0.15	0.23	0.33	A	0.013	0.011	0.013	0.014
	C	0.32	0.20	0.18	0.30	C	0.017	0.013	0.013	0.016
	G	0.27	0.26	0.20	0.28	G	0.016	0.016	0.015	0.016
	T	0.15	0.22	0.32	0.31	T	0.010	0.012	0.014	0.013
		A	C	G	T		A	C	G	T
$P^{(2)}_{G..}$	A	0.32	0.15	0.20	0.33	A	0.015	0.011	0.012	0.015
	C	0.30	0.21	0.20	0.30	C	0.014	0.013	0.013	0.015
	G	0.27	0.27	0.17	0.29	G	0.016	0.016	0.014	0.017
	T	0.22	0.20	0.22	0.36	T	0.012	0.012	0.013	0.015
		A	C	G	T		A	C	G	T
$P^{(2)}_{T..}$	A	0.35	0.16	0.14	0.35	A	0.014	0.010	0.010	0.014
	C	0.32	0.19	0.19	0.30	C	0.013	0.010	0.011	0.013
	G	0.27	0.27	0.20	0.26	G	0.012	0.012	0.011	0.012
	T	0.20	0.24	0.21	0.36	T	0.009	0.010	0.009	0.011

A Figura 6.2 mostra a probabilidade estimada de ocorrência dos segmentos homogêneos em cada posição da seqüência. Através dela, observamos que a primeira metade da seqüência, exceto no início, é constituída por um único segmento. A segunda metade do genoma, por sua vez, é constituída por subseqüências menores que se alternam entre o primeiro e o segundo segmento.

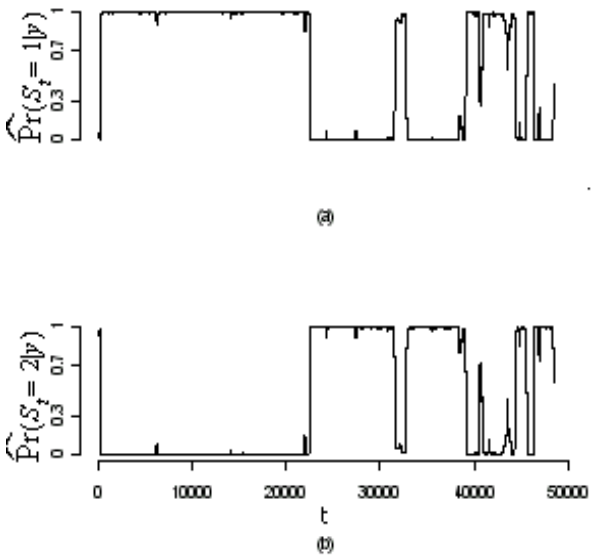


Figura 6.2: (a)-(b) Gráficos da probabilidade estimada *a posteriori* dos tipos de segmento

Capítulo 7

Considerações Finais

Neste trabalho, apresentamos os modelos Markovianos ocultos, utilizados com êxito em muitas áreas de pesquisa, principalmente em Genética e Biologia Molecular, como um caso particular das redes probabilísticas ou redes bayesianas. Estas, representadas através de um grafo, possibilitam um melhor entendimento das relações de dependência presentes em um conjunto de variáveis e a especificação de algoritmos, computacionalmente eficientes, para inferências e estimações.

Um destes algoritmos é proposto por Friedman e Koller (2003) para a estimação bayesiana da estrutura de dependência em um conjunto de variáveis a partir de dados completos, ou seja, sem existência de dados perdidos ou variáveis ocultas. Aplicamos esta metodologia em alguns conjuntos de dados simulados tomando redes de variáveis binárias e verificamos que, em alguns casos, a estimação da estrutura de dependência não é precisa e o desempenho da metodologia é insatisfatório.

A estrutura de dependência entre as variáveis, cuja estimação é mais difícil do que a estimação dos parâmetros envolvidos no modelo, é um dos elementos que caracterizam diferentes HMMs. Neste estudo, apresentamos quatro HMMs distintos, que se diferenciam através da ordem de dependência no processo aleatório oculto ou no processo observável, e discutimos métodos de estimação para cada modelo considerado.

Os dois HMMs mais simples dentre os estudados e, portanto, mais utilizados na prática são: o HMM com dependência de Markov de primeira ordem somente no processo oculto (Dugad e Desai, 1996), representado como HMM$(1, 0)$ por simplicidade de notação,

e o HMM com dependência de primeira ordem em ambos os processos aleatórios (Boys *et al.*, 2000), notado por HMM$(1,1)$. Os parâmetros envolvidos no primeiro modelo, assim como a seqüência de estados ocultos mais provável dadas as observações, são estimados através do algoritmo EM.

Os outros dois modelos propostos neste trabalho consideram dependência de segunda ordem no processo observável e de primeira ordem no oculto, representado por HMM$(1,2)$, e dependência de primeira ordem na cadeia observável e de segunda ordem na oculta, notado por HMM$(2,1)$. Para ambos os modelos são propostos métodos bayesianos de estimação.

No entanto, em situações práticas, as ordens de dependência nos processos aleatórios oculto e observado, que especificam o modelo a ser ajustado, geralmente são desconhecidas e o melhor modelo, entre os mais prováveis, precisa ser selecionado e escolhido. Para a identificação do modelo mais adequado ao conjunto de dados, propomos o uso de métodos de seleção de modelos. Tanto o fator de Bayes quanto o DIC, discutidos neste estudo para comparar o HMM$(1,1)$ e o HMM$(1,2)$, segundo as simulações realizadas, parecem ser métodos eficazes para a seleção de modelos HMM. Na aplicação desenvolvida, comparamos apenas os HMM$(1,1)$ e HMM$(1,2)$ pois, pesquisadores biológos acreditam que as ordens de dependência envolvidas nestes modelos são suficientes para modelar a estrutura presente em fragmentos das seqüências de DNA.

Aplicamos a metodologia de estimação dos modelos HMM$(1,1)$ e HMM$(1,2)$ em duas seqüências de DNA reais e verificamos, através dos métodos de seleção de modelos, o HMM mais adequado a cada conjunto de dados. A primeira seqüência analisada foi o intron 7 do gene $\alpha - fetoprotein$ dos chimpanzés, para o qual o modelo HMM$(1,1)$ com $N = 3$ segmentos homogêneos foi apontado, tanto pelo DIC quanto pelo fator de Bayes, como mais adequado do que o HMM$(1,2)$ com o mesmo número de segmentos homogêneos. Para o genoma do vírus *Bacteriophage lambda*, segunda seqüência analisada, o modelo com dependência de Markov de segunda ordem no processo aleatório observado com $N = 2$ segmentos homogêneos, HMM$(1,2)$, foi apontado, entre os seis HMMs comparados, como o mais adequado por ambos os métodos de seleção. Para as duas seqüências, apresentamos as estimativas das probabilidades de transição e das probabilidades de ocorrência dos estados ocultos em cada posição da seqüência envolvidas nos modelos selecionados.

Uma alternativa para a identificação das ordens de dependência presente em um conjunto de variáveis é o uso de metodologias, disponíveis na área de redes bayesianas, para estimação da estrutura de dependência como, por exemplo, a metodologia proposta por Friedman e Koller (2003). Assim sendo, pretendemos, em trabalhos futuros, utilizar e estender esta abordagem para a estimação dos modelos HMM, que possuem variáveis ocultas e cujos dados não são completos.

Referências Bibliográficas

[1] Aitchison, J. (1986). The statistical analysis of compositional data. London: Chapman and Hall .

[2] Bastitela, G. C. (2003). Modelos markovianos ocultos aplicados à genética. *Dissertação de Mestrado*. Universidade Federal de São Carlos, Brasil.

[3] Boys, R. e Henderson, D. (2004). A Bayesian approach to DNA Sequence Segmentation. *Biometrics* **60**:573-588.

[4] Boys, R. e Henderson, D. (2002). On determining the order of Markov dependence of an observed process governed by a hidden Markov model. *Scientifc Programming* **10**: 241-251.

[5] Boys, R., Henderson, D. e Wilkinson, D. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Applied Statistics* **49**:269-285.

[6] Braun, J. V. e Müller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science* **13**:142-162.

[7] Buntine, W. L. (1991). Theory refinement on Bayesian Networks. In: B. D. D'Ambrosio, P.Smets e P. P. Bonissone (eds.): Proc. Seventh Annual Conference on Uncertainty Artificial Intelligence (UAI'91). /San Francisco: Morgan Kaufmann, 52-60.

[8] Buntine, W. L. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering* **8**:195-210.

[9] Chib, S. e Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**(4):327-335.

[10] Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computers and Chemistry* **16**:107-115.

[11] Churchill, G. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**: 79-94.

[12] da-Silva, C. Q. (2003). Hidden Markov models applied to a subsequence of the *Xylella fastidiosa* genome. *Genetics and Molecular Biology* **26**(4):529-535.

[13] Dugad, R. e Desai, U. B. (1996). A tutorial on hidden Markov models. Technical Report No.: SPANN-96.1, Indian Institute of Technology-Bombay.

[14] Friedman, N. e Koller, D. (2003). Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* **50**:95-126.

[15] Gelman, A. e Rubin, D. B. (1992). Inference from interative simulation using multiple sequences (with discussion). *Statistical Science* **7**: 457-511.

[16] Giudici, P., Green, P. e Tarantola, C. (2000). Efficient model determination for discrete graphical models. *Biometrika*. A parecer.

[17] Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**:711-732.

[18] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time sereies and business cycle. *Econometrica* **57**:357-384.

[19] Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In: M.I. Jordan (ed.): *Learning in Graphical Models*. Dordrecht, Netherlands: Kluwer.

[20] Heckerman, D., Meek, C. e Cooper, G. (1997). A Bayesian approach to causal discovery. Technical report. Technical Report MSR-TR-97-05, Microsoft Research.

[21] Heckerman, D. e Geiger, D. (1995). Learning Bayesian Networks: a unification for discrete and Gaussian domains. In: P. Besnard e S. Hanks (eds.): *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence (UAI'95)*. San Francisco: Morgan Kaufmann, 274-284.

[22] Heckerman, D., Geiger, D. e Chickering, D. M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* **20**:197-243.

[23] Kass, R. e Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association* **90**:773-795.

[24] Madigan, D., Anderson, S., Perlman, M. e Volinsky, C. (1996). Bayesian model averaging an model selection for Markov equivalence classes of acyclic graphs. *Communications in Statistics: Theory and Methods* **25**:2493-2519.

[25] Madigan, D. e York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**:215-232.

[26] Madigan, D. e Raftery, E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal Americal Statistical Association* **89**:1535-1546.

[27] Muri, F. (1998). Modelling bacterial genomes using hidden Markov models. In *COMPSTAT'98 Proceedings in Computational Statistics*, R. W. Payne e P. J. Green (eds), 89-100. Heidelberg: Physica-Verlag.

[28] Muri, F. (1997). Comparaison d'algorithms d'identification de chaines de Markov cachées et application à la détection de regions homogènes dans les séquences d'ADN. *Phd Thesis*. Universite René Descartes, Paris.

[29] Pearl, J. (1988). Probabilistic reasoning in intelligent systems. San Francisco, Calif.: Morgan Kaufmann.

[30] Poritz, A. M. (1988). Hidden Markov models: a guided tour. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. New York: IEEE Press. **1**:7-13.

[31] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**: 257-285.

[32] Skalka, A., Burgi, E. e Hershey, A. D. (1968). Segmental distribution of nucleotides in the DNA of *Bacteriophage lambda*. *J. Molec. Biol.* **34**:1-16.

[33] Spiegelhalter, D. *et al.* (2002). Bayesian measures of model complexity and fit. *Royal Statistical Society* **64**:583-639.

[34] Zuanetti, D. A. e Milan, L. A. (2003). Análise de performance dos estimadores de modelos Markovianos ocultos. Relatório Técnico 107 do Departamento de Estatística. UFSCar.