DM

Uma Introdução às Redes Bayesianas

DISSERTAÇÃO DE MESTRADO

Joana Isabel Nunes Correia

MESTRADO EM MATEMÁTICA, ESTATÍSTICA E APLICAÇÕES

UNIVERSIDADE da MADEIRA

A Nossa Universidade

www.uma.pt

setembro | 2019

Uma Introdução às Redes Bayesianas

DISSERTAÇÃO DE MESTRADO

Joana Isabel Nunes Correia
MESTRADO EM MATEMÁTICA, ESTATÍSTICA E APLICAÇÕES

ORIENTAÇÃO
Sandra Maria Freitas Mendonça

Júri:

Doutora Ana Maria Cortesão Pais Figueira da Silva Abreu
– Professora Auxiliar da Universidade da Madeira

Doutor Sílvio Filipe Velosa
– Professor Auxiliar da Universidade da Madeira

Doutora Sandra Maria Freitas Mendonça
– Professora Auxiliar da Universidade da Madeira

Agradecimentos

Em primeiro lugar agradeço a todos os professores em geral que me acompanharam ao longo de todo o meu percurso escolar e universitário, por partilharem parte do seu conhecimento e gosto pelas várias áreas que lecionam. Agradeço por me ensinarem que há sempre mais por aprender, inspirando-me a permanecer curiosa num mundo onde há sempre mais por explorar.

Em segundo lugar tenho de agradecer aos meus colegas de mestrado pelo constante apoio e partilha de ideias ao longo destes dois anos.

Um obrigado especial à professora Sandra Mendonça, não só por propor o tema, como por ser incansável na sua orientação e pela sua dedicação a este trabalho. Esta dissertação não seria a mesma sem o seu apoio.

Por fim, o meu maior agradecimento é para os meus pais que são os meus pilares.

Resumo

As redes bayesianas são modelos de representação da realidade flexíveis por lidarem com incerteza e vários tipos de relações entre variáveis. Estas são compostas por uma parte gráfica, representada por grafos acíclicos direcionados, onde cada vértice representa uma variável aleatória, e uma parte probabilística, referente às distribuições associadas a cada um dos vértices. A estrutura do grafo define uma forma de fatorização da probabilidade conjunta das variáveis. Estas redes são úteis na inferência probabilística, facilitando o trabalho dos especialistas, permitindo diagnosticar, prever e realizar raciocínio intercausal.

Esta dissertação é constituída por sete capítulos. No primeiro é descrita a evolução dos modelos do conhecimento humano. No segundo são apresentados conceitos e definições necessários para a construção e utilização das redes bayesianas. O terceiro apresenta os métodos de inferência nestas redes, o quarto as técnicas de aprendizagem, e o quinto a análise de conflitos. No sexto são apresentados alguns comandos do programa R úteis na aplicação dos conceitos apresentados e no sétimo são apresentadas as considerações finais.

Palavras-chave: redes bayesianas, teorema de Bayes, grafos acíclicos direcionados e inferência probabilística.

Abstract

Bayesian networks are reality representing models that are flexible given that they handle uncertainty and different kinds of relationships among variables. These have a graphical aspect, given their directed acyclic graph structure, where each node represents a variable, and a probabilistic aspect, corresponding to the probabilistic distributions associated with each node. The graph structure defines a possible factorization of the joint probability of the variables. These networks are useful for probabilistic inference, helping experts in their work to perform tasks such as diagnosis, forecasts and inter-causal reasoning.

This dissertation is composed by seven chapters. The first consists of a description of the evolution of the human knowledge models. In the second, concepts and definitions necessary for the construction and use of these networks are introduced. The third presents the inference methods, the fourth presents different learning approaches and the fifth presents the conflict analysis. The sixth chapter explores some commands of R language related to bayesian concepts and inference and on seventh are presented some final considerations.

Key-words: Bayesian networks, Bayes theorem, directed acyclic graphs and probabilistic inference.

Índice

Índice de figuras

Índice de tabelas

1 Sistemas especializados

Segundo Woods *et al.* (2002) um dos principais marcos no estudo do raciocínio humano ocorreu século IV a.C., com o desenvolvimento da lógica aristotélica. Desde esta altura que se supõe existir uma relação próxima entre a lógica e o raciocínio. Segundo os especialistas da lógica as regras lógicas servem de normas ideais do raciocínio humano.

Apesar de a lógica ter tido várias etapas no seu desenvolvimento ao longo da história, Woods *et al.* (2002) argumentam que a lógica de Aristóteles foi a base do desenvolvimento do estudo do raciocínio, estudo este que inclui o estabelecimento de proposições lógicas, ainda no século III a.C., dos sistemas de dialeto formal na época medieval, do cálculo probabilístico no século XVII e do desenvolvimento da matemática moderna, estabelecida, na qual, de acordo com Woods *et al.* (2002), Frege e Peirce tiveram um papel fundamental, no século XIX. Esta última alterou o paradigma das propriedades das estruturas proposicionais do raciocínio, correspondendo a uma fase de desenvolvimento fundamental. Note-se, contudo, que a introdução da matemática na lógica foi alvo de críticas tanto por especialistas da lógica como de outras áreas.

Woods *et al.* (2002) referem ainda o facto de o raciocínio humano ser bastante suscetível de cometer falácias, e que esta disposição tende a diminuir quando esse raciocínio é aplicado num contexto prático. Deste modo, a insatisfação relativamente às limitações práticas da lógica clássica, como teoria de inferência e argumentação, e a introdução da matemática, originaram uma onda de reformas à primeira, levando ao desenvolvimento de várias subdisciplinas, como a lógica da probabilidade e a lógica computacional desde a década de 1970. As ferramentas de inteligência artificial, desenvolvidas no mesmo período, tornaram-se num aliado fundamental de aplicação dos desenvolvimentos nesta área.

Williamson (2002) salienta ainda a insuficiência da lógica dedutiva para o raciocínio, justificada pela não atribuição de grau de incerteza às premissas nem às conclusões. Assim, a lógica indutiva, suportada pelo cálculo probabilístico, surge como uma alternativa mais completa.

Relativamente à lógica computacional, Pearl (1986) define o raciocínio humano como o mecanismo pelo qual as pessoas geram conclusões ao recolherem dados de várias fontes, observando o estado do mundo, e consequentemente interpretando-o. O objetivo do seu

trabalho é a modelação computacional desse processo, combinando a probabilidade com as ciências da computação.

Segundo Kjærulff e Madsen (2008), a criação e a evolução dos computadores e o subjacente desenvolvimento da programação informática no contexto da inteligência artificial (IA)**,** tiveram desde a sua fase mais inicial a motivação de produzir sistemas capazes de modelarem o processo de raciocínio humano**,** para que fossem capazes de substituir as pessoas na resolução de tarefas repetitivas. O desenvolvimento destes sistemas evoluiu no sentido os tornar capazes de simularem a tomada de decisão, com o intuito resolverem tarefas intelectualmente desafiantes, automatizando, de acordo com Jensen (1996), a tomada de decisão em casos semelhantes.

Jensen (1996) refere o desenvolvimento duma vertente desses mesmos sistemas, a partir dos finais da década de 1960, chamada sistemas especializados. A intenção era modelar o raciocínio de especialistas de uma área particular do conhecimento, criando modelos capazes de simular eficazmente os seus processos de raciocínio, que os pudessem substituir totalmente nas suas tarefas de inferência. Era necessário então conceber um sistema capaz de modelar o conhecimento de um especialista, formalizando-o numa linguagem suportada computacionalmente, que incluísse os conceitos englobados no seu domínio do saber, e as relações entre os mesmos, tornando o sistema capaz de extrair informação a partir de dados.

Nessa altura, esses modelos tinham por base sistemas baseados em regras. De acordo com Jensen (1996), estes são constituídos por uma base de conhecimento e por um sistema de inferência. A base de conhecimento é um conjunto de regras e o sistema de inferência combina as regras com observações da realidade para extrair conclusões. As regras são formuladas da seguinte forma: se a então b. Estas têm subjacente o conceito de **causalidade,** pois a afirmação expressa uma relação determinística causal: o evento a (condição) causa o efeito b (consequência). A consequência b pode tratar-se apenas de um facto ou então de uma ação a tomar.

Uma vez que esses sistemas iniciais ignoravam os fatores de aleatoriedade e incerteza, subjacentes à resolução de problemas no mundo real, Jensen (1996) menciona uma alternativa que surgiu para lidar com essas limitações, que consistia em associar um fator de certeza a cada regra, um valor no intervalo $[-1,1]$, correspondendo o valor 1 à situação em que a conclusão é certamente verdadeira, o valor -1 à situação em que a conclusão é certamente falsa e o valor 0 à situação em que nada se pode concluir. Contudo, este mecanismo revelou-se inconsistente, tendo, segundo Kjærulff e Madsen (2008), a

inconsistência sido provada por Heckerman D. (1986), no caso das variáveis envolvidas poderem tomar mais de dois valores.

Dada a natureza incerta e subjetiva do conhecimento a partir do qual as inferências são formuladas, Pearl (1986) refere que um começo natural para expressar o raciocínio seria o seu enquadramento na teoria da probabilidade. Williamson (2002) tem a mesma visão, referindo que a probabilidade serve de representação da incerteza e deve ser incorporada no raciocínio prático, uma vez que este exige a tomada de decisões em cenários de incerteza.

No entanto, esta visão da utilidade da linguagem probabilística nem sempre vigorou na comunidade científica. Apesar de se crer, segundo Pearl (1988), que a definição de probabilidade em termos do número de formas distintas que um evento pode ocorrer foi formulada já no século XVI, essa era vista na altura apenas como um meio de comunicação de uma fonte de conhecimento determinística. Os dois séculos seguintes foram essenciais no desenvolvimento de bases axiomáticas da probabilidade na matemática, tendo surgido uma visão alternativa da probabilidade, sugerida por Bernoulli. Essa consistia na interpretação da probabilidade como "grau de confiança" que podíamos associar a um evento incerto, elemento fulcral da lógica indutiva. A aplicação da regra de Bayes, enquadrada nesta visão, foi fundamental na implementação dos processos de inferência das redes bayesianas, e que, inclusive, lhes dá o nome. A sua apresentação será feita com maior detalhe no capítulo seguinte desta dissertação. Williamson (2002) refere que apenas a partir do final do século XIX e início do século XX, com a fundamentação da teoria da probabilidade, passou a haver uma maior aceitação da probabilidade como ferramenta para a lógica. O autor refere John Keynes como um interveniente importante neste movimento, uma vez que defendia que a probabilidade generaliza a lógica através da medição do grau de conclusividade de uma afirmação. Harold Jeffreys é também referido pela sua visão da teoria da probabilidade como uma formalização da lógica dedutiva. Apenas na década de 1950 se deu o desenvolvimento da teoria de decisão estatística, proporcionando alguma aceitação na comunidade científica das redes baseadas em probabilidades.

Para além da desconfiança na utilização da probabilidade como mecanismo de medição da incerteza de ocorrência de eventos até a década de 1950, Pearl (1988) refere a hesitação inicial dos investigadores de IA em utilizar a teoria da probabilidade na modelação de domínios do conhecimento, até a década de 1970. Koller e Friedman (2009) ressaltam o custo computacional inicial elevado, em termos da capacidade de memória de armazenamento, mesmo para um número relativamente pequeno de variáveis, e a dificuldade na obtenção de valores razoáveis para as probabilidades associadas a estas variáveis por parte

de especialistas. Pearl (1988) vai mais longe e refere que os cientistas consideravam que ainda que as pessoas não fossem boas estimadoras de probabilidades, o uso das probabilidades estimadas pelas mesmas seria inconveniente, dado o custo computacional envolvido. Estes fatores apresentavam-se então como barreiras à utilização do cálculo probabilístico na automação da inferência, que apenas foram superados através do desenvolvimento de metodologias de representação das relações de dependência entre as variáveis, nomeadamente através de grafos, e a aplicação de teoremas de cálculo probabilístico.

Quanto à incapacidade dos humanos fornecerem boas estimativas de probabilidades, Pearl (1988) refere que as pessoas são melhores a conceber relações qualitativas do que estimativas absolutas. Assim, apesar dos valores estimados poderem estar completamente errados, desde que os rácios dos números atribuídos às variáveis, que compõem um dado problema, reflitam corretamente as suas relações, as conclusões serão consistentes. Pearl (1988) refere ainda que a estimação probabilística permite modelar compactamente o conhecimento, fornecendo um mecanismo de manipulação do conhecimento modelado.

Houve então uma mudança de paradigma a partir da década de 1970, quando os investigadores voltaram a sua atenção para uma interpretação probabilística dos fatores de incerteza. Esta opção revelou-se como uma forma mais segura e prática de lidar com a incerteza, introduzindo medidas quantitativas da mesma.

Neste sentido, o uso de sistemas especializados baseados em regras foi substituído pelo uso de os sistemas especializados normativos. Segundo Jensen (1996), estes têm a particularidade de basearem o seu raciocínio na teoria da decisão e no **cálculo probabilístico clássico**, englobando o efeito da incerteza, ao invés dos cálculos de incerteza incoerentes anteriormente utilizados. Como referem Kjærulff e Madsen (2008), o termo "normativo" advém do facto do comportamento destes sistemas ser orientado por um conjunto de regras fundamentais ou axiomas, essenciais para a capacidade de inferência destes sistemas, pois permitem ultrapassar as barreiras a nível armazenamento computacional mencionadas por Koller e Friedman (2009). É também referido por Jensen (1996) que, ao contrário dos anteriores, estes sistemas modelam o domínio do conhecimento e não o especialista, servindo de apoio na sua tarefa de inferência, ao invés de o substituir.

Uma vez provada a eficaz capacidade dos **grafos** na representação de relações de dependência e independência entre variáveis, estes passaram a ser a linguagem de eleição para modelar domínios do conhecimento, sendo essenciais para a aceitação da teoria da probabilidade, anteriormente evitada pelos investigadores de IA. Assim, Kjærulff e Madsen

(2008) mencionam que o conceito de sistemas especializados passou a surgir geralmente associado ao de **redes probabilísticas**. Estas correspondem a modelos gráficos nos quais os vértices representam as variáveis e as arestas direcionadas (ou orientadas) representam as interações entre as mesmas. Um aspeto fulcral destas que as distingue de outros sistemas de inferência automatizada é a sua possibilidade de **raciocínio intercausal**. Kjærulff e Madsen (2008) define este como o processo pelo qual a obtenção de evidência acerca duma hipótese leva à diminuição automática na crença das hipóteses concorrentes, não observadas, constituindo, segundo Pearl (1988) um mecanismo de inferência seguro e completo.

As **redes bayesianas**, tema central desta dissertação, podem ser vistas, segundo Kjærulff e Madsen (2008), como uma extensão dos modelos iniciais de representação e manipulação de conhecimento, incluem-se na família dos sistemas especializados normativos e das redes probabilísticas.

Segundo Williamson (2002), e apesar de muitas abordagens lógicas da teoria da probabilidade falharem de um ponto de vista filosófico, as redes bayesianas oferecem uma solução para a resolução de problemas práticos que lidam com a lógica que incorpora conceitos probabilísticos. O autor refere vantagens técnicas destas redes relativamente a outros modelos, como o facto de não terem problemas de consistência, do número de probabilidades necessárias especificar ser relativamente pequeno e da estimação dessas mesmas probabilidades ser relativamente rápida, consoante a dimensão da rede e as técnicas de propagação aplicadas. Também por serem poucos os modelos que lidam com incerteza, os autores referem que as redes bayesianas constituem uma oportunidade prática para tal. Segundo estes, a sua utilidade vai para além do formalismo técnico que fornecem na estruturação do conhecimento, permitindo qualquer tipo de raciocínio com variáveis causais, tal como raciocínios de diagnóstico, previsão ou explicação causal.

Note-se que a famosa afirmação de George E. P. Box segundo a qual “todos os modelos estão errados, mas alguns são úteis” aplica-se também aos modelos probabilísticos em geral, e às redes bayesianas em particular. Neste sentido, Kjærulff e Madsen (2008) ressaltam que, como qualquer modelo, os modelos probabilísticos não representam a realidade na sua totalidade, incluindo unicamente mecanismos relevantes da área do conhecimento da qual se pretende inferir conclusões. Esta noção é também referida por Pearl (1988). Segundo este autor, a modelação realista do raciocínio, no contexto de um domínio do conhecimento, requer simplificações na especificação do mesmo. Essas simplificações podem passar por não referir alguns factos ou então por referi-los de forma bastante sumarizada. Em

termos das redes bayesianas a ocultação de factos traduz-se na não introdução de algumas variáveis e/ou interações entre variáveis.

2 Apresentação das redes bayesianas

2.1 Apresentação

Usadas como meio de modelação de domínios do conhecimento, Kjærulff e Madsen (2008) descrevem resumidamente as redes bayesianas como grafos acíclicos direcionados que definem a fatorização da probabilidade de um conjunto de variáveis aleatórias, que compõem o modelo. As variáveis aleatórias que compõem uma rede bayesiana são representadas pelos vértices do grafo, tendo cada uma associada uma distribuição probabilística. Os vértices, correspondentes às variáveis aleatórias, são ligados por arestas direcionadas. Estas ligações representam relações de dependência entre as variáveis, que indicam também relações de independência condicional de outras variáveis. O nome *redes bayesianas* advém do papel central no processo de inferência das mesmas que toma a regra de Bayes, de Thomas Bayes (1702-1761).

Da apresentação informal anterior de rede bayesiana, e de acordo com Kjærulff e Madsen (2008), conclui-se que a estas redes estão associados dois aspetos, nomeadamente o qualitativo, relacionado com a estrutura gráfica, e o quantitativo, relativo à parte probabilística e numérica. Ambos se sustentam mutuamente e são fundamentais na definição, construção e subjacente processo de inferência dos modelos bayesianos. Assim, apresentam-se os conceitos básicos associados aos aspetos qualitativo e quantitativo, necessários à compreensão dos mecanismos de construção e de utilização das redes bayesianas como ferramenta de inferência.

2.2 Aspeto quantitativo

Tratando-se de um modelo probabilístico o formalismo associado às redes bayesianas consiste na estimação de probabilidades de variáveis da rede. Este processo baseia-se nos parâmetros numéricos, presentes nas tabelas de probabilidade, associadas a cada vértice do grafo que representa a rede em questão, e na subsequente manipulação desses parâmetros de acordo com a linguagem probabilística.

A inferência em cenários de incerteza permitida pelas redes bayesianas traduz-se em termos práticos na manipulação de probabilidades condicionais, para raciocinar sobre variáveis cujo estado não foi observado, estimando a sua probabilidade, dada evidência, ou seja, tendo em conta a observação do estado de outras variáveis. Importa então apresentar as definições básicas da teoria da probabilidade que suportam o funcionamento destas redes.

Comecemos por definir formalmente o **conceito de probabilidade**. Consideremos Ω o conjunto dos resultados de uma experiência aleatória, *i.e.*, uma experiência cujo resultado é desconhecido, mas cujo conjunto dos resultados possíveis é conhecido. Dizemos que $\mathscr{A}$ é uma sigma-álgebra de acontecimentos de Ω se satisfizer as seguintes propriedades:

i. $\Omega \in \mathscr{A}$;

ii. Se $A \in \mathscr{A}$, então $\Omega - A \in \mathscr{A}$;

iii. Se $A_1, A_2, \ldots \in \mathscr{A}$, então $\bigcup_{i=1}^{\infty} A_i \in \mathscr{A}$.

Segundo Kolmogorov (1950), uma probabilidade é uma aplicação P: $\mathscr{A} \rightarrow [0,1]$ tal que:

- $P(\Omega)=1$;
- $P(A) \geq 0$, para qualquer acontecimento $A \in \mathscr{A}$;
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$, quando A_1, A_2,... são acontecimentos mutuamente exclusivos, dois a dois.

Como a linguagem probabilística é constituída por proposições acerca da probabilidade de eventos, importa referir que um evento ou acontecimento pode ser considerado como o resultado de uma experiência, pode ser uma observação particular do valor de uma variável aleatória ou resultar da atribuição de um valor a uma variável aleatória. As probabilidades associadas a uma variável aleatória discreta X representam-se por $P(X)$. Nesta dissertação serão consideradas apenas variáveis aleatórias discretas que tomam um número finito de estados. Se definirmos o seu **suporte** como sendo o conjunto dos seus estados, $sup(X) = \{x_1, \ldots, x_n\}$, $n \in \mathbb{N}$, então $P(X) = (P(X = x_1), \ldots, P(X = x_n))$ representa as probabilidades associadas a cada estado, que pode ser denotada abreviadamente por $P(X) = \big(P(x_1), \ldots, P(x_n)\big)$.

Williamson (2002) refere que o ponto de partida da maioria das teorias que integram probabilidade e lógica baseia-se na associação de probabilidades a fórmulas proposicionais e apresenta uma forma de fazer esta associação definindo uma probabilidade sobre o conjunto das fórmulas proposicionais dada uma linguagem proposicional.

Assume-se, nos exemplos apresentados ao longo do texto desta secção, que as variáveis são binárias, tomando apenas dois estados, que no caso de uma variável aleatória X, que designaremos simplesmente por variável X, podem identificar-se por $X = x_1$ e $X = x_2$. A distribuição desta variável define-se então por $P(X) = (P(x_1), P(x_2))$. Esta distribuição pode,

por exemplo, tomar os valores indicados na seguinte tabela. Note-se que a soma das probabilidades deve ser igual a 1.

Tabela 2.1 – Exemplo de distribuição $P(X)$.

X	$P(X)$
x_1	0.61
x_2	0.39
$Total$	1

Definida a distribuição de probabilidade de uma variável X, importa definir a distribuição conjunta de duas variáveis, por exemplo X e Y, denotada por $P(X,Y)$, tendo cada variável como respetivo suporte, digamos, $sup(X) = \{x_1, \dots, x_n\}$ e $sup(Y) = \{y_1, \dots, y_m\}$, $n, m \in \mathbb{N}$. Assumindo novamente que as variáveis são binárias, e que tomam os valores dos espaços de estados $\{x_1, x_2\}$ e $\{y_1, y_2\}$, então $P(X,Y) = \big(P(x_1,y_1), P(x_1,y_2), P(x_2,y_1), P(x_2,y_2)\big)$. Considerando os valores da distribuição anterior $P(X)$ da variável X, as distribuições de $P(Y)$ e $P(X,Y)$ podem, a título de exemplo, tomar os valores especificados nas seguintes tabelas:

Tabela 2.2 – Exemplo de distribuição $P(Y)$.

Y	$P(Y)$
y_1	0.82
y_2	0.18
$Total$	1

Tabela 2.3 – Exemplo de distribuição conjunta $P(X,Y)$.

X	Y	$P(X,Y)$
x_1	y_1	0.54
x_1	y_2	0.07
x_2	y_1	0.28
x_2	y_2	0.11
$Total$		1

Note-se que a partir da distribuição $P(X,Y)$ se podem obter as distribuições individuais $P(X)$ e $P(Y)$. A operação de eliminar uma variável de uma distribuição conjunta chama-se **marginalização**, e toma um papel importante na inferência bayesiana. No caso de uma distribuição conjunta de duas variáveis, depois de escolhida a variável a eliminar, o processo de marginalização dessa variável passa por somar separadamente todas as entradas, da tabela de distribuição conjunta, referentes a cada um dos estados iguais da variável que se deseja manter. Depois forma-se a distribuição dessa variável não marginalizada, através da junção de todos os valores associados à probabilidade de cada um desses possíveis estados. Este é resultado da aplicação do conhecido t**eorema da probabilidade total**. Especificando, por exemplo, que desejamos obter $P(X)$ a partir de uma distribuição $P(X,Y)$, somam-se separadamente todas as entradas da tabela $P(X,Y)$ referentes a cada x_i, obtendo-se a probabilidade associada a cada um desses n estados x_i de X. Algebricamente podemos escrever a probabilidade de cada estado x_i de X como:

$$P(x_i) = \sum_{j=1}^{m} P(x_i, y_j) = P(x_i, y_1) + \cdots + P(x_i, y_m), \forall i \in \{1, \ldots, n\}. \quad (2.1)$$

Uma vez calculadas as probabilidades associadas a cada um dos n estados de X, obtemos então a distribuição total da variável X, $P(X)$, composta por esses mesmos valores:

$$P(X) = \left(\sum_{j=1}^{m} P(x_1, y_j), \ldots, \sum_{j=1}^{m} P(x_n, y_j)\right). \quad (2.2)$$

A partir do exemplo numérico apresentado, calculamos a probabilidade de cada estado x_i:

$$P(X = x_1) = \sum_{j=1}^{2} P(x_1, y_j) = P(X = x_1, Y = y_1) + P(X = x_1, Y = y_2) = 0.54 + 0.07 = 0.61, \quad (2.3)$$

$$P(X = x_2) = \sum_{j=1}^{2} P(x_2, y_j) = P(X = x_2, Y = y_1) + P(X = x_2, Y = y_2) = 0.28 + 0.11 = 0.39. \quad (2.4)$$

Após a marginalização da variável Y, da distribuição $P(X, Y)$, obtemos a distribuição original de X:

$$P(X) = P(X = x_1, X = x_2) = (0{,}61; 0{,}39). \quad (2.5)$$

Note-se que um modelo pode conter mais de duas variáveis. No caso das redes bayesianas, a probabilidade conjunta fatorizada por essas, contêm tantas variáveis quanto aquelas que compõem o modelo representado pela rede. Nesse caso, para o processo de inferência, pode ser necessário marginalizar mais do que uma variável da distribuição conjunta completa. Suponhamos que ao exemplo considerado se adiciona uma nova variável Z, com $P(Z) = (P(z_1), P(z_2))$, sendo a distribuição $P(X, Y, Z)$ apresentada na tabela:

Tabela 2.4 – Exemplo de distribuição conjunta $P(X,Y,Z)$.

X	Y	Z	$P(X,Y,Z)$
x_1	y_1	z_1	0,26
x_1	y_1	z_2	0,28
x_1	y_2	z_1	0,03
x_1	y_2	z_2	0,04
x_2	y_1	z_1	0,19
x_2	y_1	z_2	0,09
x_2	y_2	z_1	0,05
x_2	y_2	z_2	0,06
$Total$			1

A partir da distribuição podem-se obter as distribuições individuais de cada variável, nomeadamente $P(X)$, $P(Y)$, $P(Z)$, ou então as distribuições conjuntas $P(X, Y)$, $P(X, Z)$ e

$P(Y,Z)$. Se desejamos obter, por exemplo $P(X,Y)$, teremos de marginalizar Z, somando todas as entradas consistentes para cada par (x_i, y_j). Supondo $\sup(Z) = \{z_1, \dots, z_l\}, (l \in \mathbb{N})$, essa operação apresenta-se algebricamente por:

$$P(x_i, y_j) = \sum_{k=1}^{l} P(x_i, y_j, z_k) = P(x_i, y_j, z_1) + \cdots + P(x_i, y_j, z_l), \forall i \in \{1, \dots, n\}, \forall j, \in \{1, \dots, m\}. \quad (2.6)$$

Aplicando essa operação ao exemplo numérico, temos:

$$P(X = x_1, Y = y_1) = \sum_{k=1}^{2} P(x_1, y_1 z_k) = 0.26 + 0.28 = 0.54, \quad (2.7)$$

$$P(X = x_1, Y = y_2) = \sum_{k=1}^{2} P(x_1, y_{2,} z_k) = 0.03 + 0.04 = 0.07, \quad (2.8)$$

$$P(X = x_2, Y = y_1) = \sum_{k=1}^{2} P(x_2, y_{1,} z_k) = 0.19 + 0.09 = 0.28, \quad (2.9)$$

$$P(X = x_2, Y = y_2) = \sum_{k=1}^{2} P(x_2, y_{2,} z_k) = 0.05 + 0.06 = 0.11. \quad (2.10)$$

Podemos então escrever:

$$P(X,Y) = \big(P(x_1, y_1), P(x_1, y_2), P(x_2, y_1), P(x_2, y_2)\big) = (0{,}54; 0{,}07; 0{,}28; 0{,}11), \quad (2.11)$$

valores que coincidem com os indicados na **Erro! A origem da referência não foi encontrada.** r eferente a $P(X,Y)$. Uma vez obtida a distribuição de $P(X,Y)$, a partir de $P(X,Y,Z)$ marginalizando Z, para obter $P(X)$ basta marginalizar Y dessa distribuição, seguindo o processo já descrito.

Uma vez que a inferência bayesiana se baseia no pressuposto da observação do estado de uma ou mais variáveis, uma operação recorrente na inferência bayesiana é o **condicionamento probabilístico** da distribuição conjunta de variáveis, a um estado particular de uma ou mais **variáveis aleatórias**. Baseando-nos no exemplo anterior da distribuição de $P(X,Y)$, se assumirmos a observação do estado da variável $Y = y_1$, deve-se condicionar a distribuição $P(X,Y)$ a essa observação, obtendo a distribuição $P(X,Y|Y = y_1)$. Note-se que em termos de modelos probabilísticos, $P(X)$ representa o conhecimento acerca da variável X, antes de se ter algum conhecimento sobre a variável Y. Para essa operação deve-se começar por selecionar as entradas da primeira tabela $P(X,Y)$ consistentes com $Y = y_1$, obtendo uma tabela menor:

Tabela 2.5 – Entradas do exemplo de distribuição conjunta $P(X,Y)$ com $Y=y_1$.

X	Y	$P(X,Y)$
x_1	y_1	0.54
x_2	y_1	0.28
$Total$		0.82

Note-se que a soma total das probabilidades é diferente de 1. Para alterar essa condição, obtendo uma distribuição de probabilidade, deve-se proceder à operação de **normalização** desta distribuição, dividindo cada probabilidade pela soma atual total:

Tabela 2.6 – Normalização da distribuição $P(X,Y)$ com $Y=y_1$.

X	Y	$P(X,Y\|Y=y_1)$
x_1	y_1	$0.54/0{,}82 \approx 0{,}66$
x_2	y_1	$0.28/0{,}82 \approx 0{,}34$
$Total$		1

Esta operação de condicionar uma distribuição conjunta ao estado de uma ou mais das suas variáveis, é importante no que toca à inferência bayesiana, uma vez que, como referido na definição de rede bayesiana, a cada vértice é associada uma distribuição de **probabilidade condicional**. Deste modo, dadas duas variáveis X e Y, temos que probabilidade da variável X condicionada à variável Y denota-se por $P(X|Y)$, que expressa a probabilidade associada a cada estado de X, dado que se observou um certo estado de Y. Abreviadamente escreve-se a distribuição da variável X condicionada a Y por:

$$P(X|Y) = \big(P(X = x_1|Y = y_1), \dots, P(X = x_n|Y = y_m)\big). \qquad (2.12)$$

Em termos práticos a operação necessária para a especificação desta distribuição consiste em estender a operação anterior, ao condicionamento a todos possíveis estados de Y. Neste sentido, no exemplo, para se obter $P(X|Y)$ basta estimar $P(X|Y = y_2)$, e juntar ao condicionamento $P(X|Y = y_1)$, já estimado. Temos então essa operação apresentada na seguinte tabela, onde se pode notar que a soma de cada entrada referente a cada estado da variável que condiciona tem valor igual a 1, como esperado:

Tabela 2.7 – Distribuição condicionada $P(X|Y)$ obtida através da distribuição conjunta $P(X, Y)$.

X	Y	$P(X\|Y)$	$Total$
x_1	y_1	$0{,}54/0{,}82 \approx 0.66$	1
x_2		$0{,}28/0{,}82 \approx 0{,}34$	
x_1	y_2	$0{,}07/0{,}18 \approx 0{,}39$	1
x_2		$0{,}11/0{,}18 \approx 0.61$	

Dados os cálculos apresentados na tabela anterior podemos escrever uma expressão geral que permite obter a distribuição $P(X|Y)$ dados os respetivos suportes de X e Y, calculando[1]:

$$\forall i,j, P(x_i|y_j) = \frac{P(x_i, y_j)}{P(y_j)}. \quad (2.13)$$

Note-se que a probabilidade condicional não está definida quando $P(y_j) = 0$. No entanto como $y_j \in sup(Y)$, temos que $P(y_j) > 0$, $\forall i,j$.

Em alternativa podemos escrever:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}. \quad (2.14)$$

A **regra de Bayes** encontra-se intrinsecamente relacionada com a anterior definição, por resultar de uma manipulação da mesma. Esta regra toma um papel central nos cálculos da inferência bayesiana, ajudando a simplificar os cálculos intermédios. Segundo esta:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}. \quad (2.15)$$

É importante referir que apesar do exemplo contemplado apenas apresentar uma variável condicionada a outra, não há, novamente, um limite no número de variáveis condicionadas nem das que condicionam uma distribuição de probabilidade condicional.

Note-se que no contexto bayesiano se atribuem designações às componentes da igualdade (2.15), nomeadamente a distribuição $P(X|Y)$ é designada de distribuição *a posteriori*, a distribuição $P(X)$ de distribuição *a priori*. A probabilidade condicional $P(X|Y)$ pode também denotar-se por $L(Y|X)$, que se lê como "a **verosimilhança** (*likelihood*) de Y dado X". Deste modo, considerando que se sabe o estado da variável X, e que $\{y_1, \dots, y_n\}$ constitui o conjunto dos possíveis estados da variável Y que têm efeito em X, então, segundo Jensen (1996), $P(X|Y_i)$ define uma medida de quão provável que é $Y = y_i$ ser a causa de X.

Como referido, uma propriedade probabilística central na ótica bayesiana é a dependência entre variáveis. Contudo, em termos de distribuição probabilística é a relação oposta, ou seja, a de **independência entre variáveis,** que tem implicações matemáticas úteis em termos da definição de distribuição conjunta. Assim, dadas duas variáveis X e Y, dizemos

[1] Por comodidade de escrita, e sempre que tal não cause confusão, omitiremos os suportes das variáveis.

que estas são independentes, e denotamos essa relação por $X \perp Y$, se para cada um dos seus estados, respetivamente x_i e y_i, temos:

$$P(x_i, y_j) = P(x_i)P(y_j), \forall i, j. \tag{2.16}$$

Podemos escrever de forma simplificada:

$$P(X, Y) = P(X)P(Y). \tag{2.17}$$

Se essa relação se verificar, então podemos desenvolver algebricamente a definição de probabilidade condicional:

$$\forall i, j, P(x_i|y_j) = \frac{P(x_i, y_j)}{P(y_j)} = \frac{P(x_i)P(y_j)}{P(y_j)} = P(x_i). \tag{2.18}$$

Conclui-se deste modo que é inútil condicionar a probabilidade de uma variável a outra quando essas são independentes. Podemos definir a relação de independência $X \perp Y$ pela seguinte expressão:

$$\forall i, j: P(x_i|y_j) = P(x_i), \tag{2.19}$$

ou então, de forma sintetizada:

$$P(X|Y) = P(X). \tag{2.20}$$

Existem ainda relações de **independência condicional** que definem a relação de independência de duas variáveis condicionada à observação de uma terceira, ou ainda de um subconjunto de outras variáveis. Dadas as variáveis X, Y e Z, dizemos que as variáveis aleatórias X e Y são independentes dada a variável Z, e escrevemos $X \perp Y\ |Z$, se, e só se:

$$\forall i, j, k: P(x_i, y_j|z_k) = P(x_i|z_k)P(y_j|z_k), \tag{2.21}$$

ou, equivalentemente,

$$\forall i, j, k: P(x_i|y_{j,}z_k) = P(x_i|z_k), \tag{2.22}$$

ou ainda, equivalentemente:

$$\forall i, j, k: P(y_{j,}|x_i, z_k) = P(y_{j,}|z_k). \tag{2.23}$$

A expressão (2.21) advém da adaptação, à existência de uma variável condicionante, da definição de independência apresentada pela expressão (2.16), enquanto que as

expressões (2.22) e (2.23) advêm da adaptação da expressão (2.19) que define a distribuição de probabilidade de uma variável condicionada a outra.

Por exemplo, a implicação (2.21) ⇒ (2.22) pode ser provada da seguinte forma:

$$\begin{aligned} \forall i,j,k,: P(x_i|y_j,z_k) &= \frac{P(x_i,y_j,z_k)}{P(y_j,z_k)} \\ &= \frac{P(x_i,y_j|z_k)P(z_k)}{P(y_j,|z_k)P(z_k)} \\ &= \frac{P(x_i|z_k)P(y_j|z_k)}{P(y_j,|z_k)} \\ &= P(x_i|z_k). \end{aligned} \tag{2.24}$$

Uma regra importante no que toca à fatorização das variáveis é a **regra da cadeia**. Esta define a distribuição de um conjunto de variáveis $(X_1,\dots,X_n)$ pela multiplicação sucessiva de probabilidades condicionadas e conjuntas dessas mesmas variáveis, do seguinte modo:

$$\begin{aligned} P(X_1,\dots,X_n) &= P(X_1|X_2,\dots,X_n)P(X_2,\dots,X_n) \\ &= P(X_1|X_2,\dots,X_n)P(X_2|X_3,\dots,X_n)P(X_3,\dots,X_n) \\ &\dots \\ &= P(X_1|X_2,\dots,X_n)P(X_2|X_3,\dots,X_n)\dots P(X_{n-1}|X_n)P(X_n) \\ &= \left[\prod_{i=1}^{n-1} P(X_i|X_{i+1},\dots,X_n)\right] P(X_n). \end{aligned} \tag{2.25}$$

Conclui-se pelo desenvolvimento da expressão anterior que, para cada conjunto de n variáveis, a regra da cadeia permite definir $n!$ formas de fatorizar a distribuição probabilística conjunta dessas mesmas variáveis, através da multiplicação de probabilidades condicionadas e marginais. Note-se que para um conjunto de n variáveis binárias, a especificação da respetiva probabilidade conjunta exige a especificação de uma tabela com 2^n entradas. Assim, o acréscimo do número de variáveis resulta na exigência de especificação de tabelas de dimensões cada vez maiores para representar essa distribuição. Deste modo, a definição da regra da cadeia permite a simplificação do processo, possibilitando que essa distribuição conjunta possa ser repartida em várias distribuições de menores dimensões.

Por exemplo, considerando o conjunto de variáveis binárias X, Y e Z e a correspondente distribuição probabilística conjunta $P(X,Y,Z)$, de acordo com a regra de cadeia podemos escrever as seis possíveis factorizações seguintes:

$$
\begin{aligned}
P(X,Y,Z) &= P(X|Y,Z)P(Y|Z)P(Z) && (2.26)\\
&= P(X|Y,Z)P(Z|Y)P(Y) && (2.27)\\
&= P(Y|X,Z)P(X|Z)P(Z) && (2.28)\\
&= P(Y|X,Z)P(Z|X)P(X) && (2.29)\\
&= P(Z|X,Y)P(X|Y)P(Y) && (2.30)\\
&= P(Z|X,Y)P(Y|X)P(X). && (2.31)
\end{aligned}
$$

Constatamos então, que a especificação de uma tabela de distribuição conjunta pode ser reduzida à especificação de outras tabelas menores, que multiplicadas constituem a distribuição conjunta. Há também que referir que para além de ser vantajoso reduzir uma distribuição conjunta na multiplicação de outras menores, uma vez que nessas se incluem probabilidades condicionadas, será vantajoso escolher probabilidades condicionadas com as menores dimensões possíveis. Por exemplo, uma distribuição $P(X|Y_1,\dots,Y_n)$ exige a especificação de uma tabela com 2×2^n entradas. Assim a diminuição do número das variáveis que condicionam nessas probabilidades mostra-se vantajosa por permitir a diminuição da tabela exigida.

Regressando à distribuição conjunta das variáveis X, Y e Z, se assumirmos que $X \perp Y|Z$, e dadas as definições de independência condicional, diminui o número de fatorizações possíveis da distribuição conjunta dessas variáveis $P(X,Y,Z)$ definidas pela regra da cadeia. Observamos que, por exemplo, as expressões (2.26) e (2.28) ficam reduzidas a:

$$P(X,Y,Z) = P(X|Z)P(Y|Z)P(Z). \qquad (2.32)$$

Esta distribuição é composta por uma distribuição de probabilidade individual, $P(Z)$, e duas condicionais, $P(X|Z)$ e $P(Y|Z)$, sendo apenas necessário especificar uma tabela com 2 entradas e outras duas tabelas com 4 entradas $(2\times 2^1 = 4)$. Assim, considerando as probabilidades condicionadas, cujo condicionamento está subjacente a apenas a uma variável, consegue-se obter distribuições de dimensões menores, comparativamente a outras fatorizações, especificadas pela regra da cadeia, compostas por probabilidades condicionais com mais do que uma variável condicionante.

Na secção seguinte é apresentada a forma de leitura destas relações a partir dos grafos que representam as redes bayesianas.

Terminaremos esta secção com a apresentação da distribuição de Dirichlet, necessária para a definição de uma medida de *score* na subsecção 4.1.2. Dado um número inteiro positivo n e um vetor aleatório absolutamente contínuo $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$, dizemos que $\boldsymbol{X}$ tem

distribuição de Dirichlet, com vetor de parâmetros positivos $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$, e escrevemos $\boldsymbol{X} \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_n)$, se $\boldsymbol{X}$ tiver função densidade conjunta dada por:

$$f_X(\boldsymbol{x}) = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)} \left(\prod_{i=1}^{n} x_i^{\alpha_i - 1} \right), \tag{2.33}$$

para $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$, com $x_i \in [0,1]$, $\forall i \in \{1, .., n\}$, tais que $\sum_{i=1}^{n} x_i = 1$. Na expressão, Γ denota a função gama, que se encontra relacionada com a função beta, B, através da expressão[2]:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}, \text{para } \alpha, \beta > 0. \tag{2.34}$$

Mais informações sobre a distribuição de Dirichlet podem ser encontradas, por exemplo, em Kotz *et al*. (2019).

Uma das propriedades que faz a distribuição de Dirichlet especial é o facto de ser a distribuição *a priori* conjugada da distribuição multinomial, o que simplifica os cálculos, em algumas situações. Isto significa que se um vetor aleatório (de probabilidades, por exemplo) $\boldsymbol{P} = (P_1, \dots, P_n)$ tiver (*a priori*) uma distribuição de Dirichlet com parâmetros $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$, *i.e.*, se, tomando $\boldsymbol{p} = (p_1, \dots, p_n)$, a função densidade de probabilidade for dada por:

$$f_{\boldsymbol{P}}(\boldsymbol{p}) = \frac{\prod_{i=1}^{n} \Gamma(a_i)}{\Gamma\left(\sum_{i=1}^{n} a_i\right)} \left(\prod_{i=1}^{n} p_i^{a_i - 1} \right) \propto \prod_{i=1}^{n} p_i^{a_i - 1}, \tag{2.35}$$

onde o símbolo $\propto$ expressa proporcionalidade, $p_i \in [0,1]$, $\forall i \in \{1, .., n\}$ e $\sum_{i=1}^{n} p_i = 1$, se os dados, $x = (x_1, \dots, x_n)$ provenientes de um vetor aleatório $\boldsymbol{X} = (X_1, \dots, X_n)$, tiverem uma

[2] Como a relação $\sum_{i=1}^{n} x_i = 1$ implica que conhecidas, por exemplo, as $n - 1$ primeiras variáveis, $x_1, x_2, \dots, x_{n-1}$, a última tem um valor conhecido, $x_n = 1 - \sum_{i=1}^{n-1} x_i$, a definição da distribuição de Dirichlet anterior é por vezes apresentada como sendo a distribuição do vetor aleatório $\boldsymbol{X}' = (X_1, X_2, \dots, X_{n-1})$, sendo a sua função densidade de probabilidade dada por:

$$f_{\boldsymbol{X}'}(\boldsymbol{x}') = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)} \left(\prod_{i=1}^{n-1} x_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{n-1} x_i \right)^{\alpha_n - 1},$$

para $\boldsymbol{x}' = (x_1, x_2, \dots, x_{n-1})$, com $x_i \in [0,1]$, $\forall i \in \{1, .., n-1\}$, tais que $\sum_{i=1}^{n-1} x_i \leq 1$. Nesta forma, facilmente se constata que a distribuição de Dirichlet é uma possível generalização ao caso multivariado da distribuição beta já que para $n = 2$, caso em que o vetor $\boldsymbol{X}'$ fica reduzido à variável X_1, obtemos a função densidade de probabilidade (tomando $x = x_1$),

$$f_{X_1}(x) = \frac{1}{B(\alpha_1, \alpha_2)} x^{\alpha_1 - 1} (1 - x)^{\alpha_2 - 1}, x \in [0,1],$$

i.e., a função densidade de probabilidade da distribuição Beta(α_1, α_2).

distribuição multinomial de parâmetros $(m, p_1, \dots, p_n)$, onde m é um número inteiro positivo, *i.e.*, se

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{P} = \boldsymbol{p}) = \frac{m!}{x_1! \dots x_n!} \prod_{i=1}^{n} p_i^{x_i}, \tag{2.36}$$

com $\boldsymbol{x} \in \mathbb{N}^n$ e $\sum_{i=1}^{n} x_i = m$, então a distribuição *a posteriori* de $\boldsymbol{P}$:

$$\begin{aligned} f_{\boldsymbol{P}|\boldsymbol{X}=\boldsymbol{x}}(\boldsymbol{p}) &= \frac{P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{P} = \boldsymbol{p}) f_{\boldsymbol{P}}(\boldsymbol{p})}{P(\boldsymbol{X} = \boldsymbol{x})} \propto P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{P} = \boldsymbol{p}) f_{\boldsymbol{P}}(\boldsymbol{p}) \\ &\propto \left(\frac{m!}{x_1! \dots x_n!} \prod_{i=1}^{n} p_i^{x_i} \right) \left(\prod_{i=1}^{n} p_i^{a_i - 1} \right) \propto \prod_{i=1}^{n} p_i^{x_i + a_i - 1}, \end{aligned} \tag{2.37}$$

i.e., é uma distribuição de Dirichlet com parâmetros $\boldsymbol{\alpha} = (x_1 + \alpha_1, \dots, x_n + \alpha_n)$.

Mais geralmente, sendo $\boldsymbol{\theta}$ o conjunto dos parâmetros desconhecidos e $\boldsymbol{x}$ a amostra aleatória em causa, se a distribuição *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{x})$ e a distribuição *a priori* $p(\boldsymbol{\theta})$ forem da mesma família de distribuições de probabilidade, dizemos que as duas distribuições são conjugadas, e que a distribuição *a priori* é a conjugada *a priori* para a função de verosimilhança $p(\boldsymbol{x}|\boldsymbol{\theta})$. Mais informações sobre este assunto podem ser encontradas, por exemplo, em Raiffa e Schlaifer (1961).

2.3 Aspeto qualitativo

Como referido no início do presente capítulo, e ao longo da seção anterior (2.2), os modelos probabilísticos, nomeadamente as redes bayesianas, são representados através de grafos. Estas estruturas de dados servem de representação das distribuições probabilísticas. Consequentemente, torna-se importante apresentar alguns conceitos básicos relacionados com estas estruturas.

Um **grafo** pode ser definido como um par $G = (V, A)$, onde V é um conjunto finito de vértices distintos, e A é um conjunto de arestas, que ligam vértices, com $A \subseteq V^2$. Para indicar a ligação de um vértice v_1 a um vértice v_2 por uma aresta direcionada usa-se a notação $v_1 \rightarrow v_2$. Se a aresta não for direcionada, essa ligação denota-se por $v_1 - v_2$. Para representar qualquer tipo de ligação entre dois vértices v_1 e v_2, nomeadamente $v_1 - v_2$, $v_1 \rightarrow v_2$ ou $v_1 \leftarrow v_2$, usamos a notação $v_1 \sim v_2$. Um par de vértices ligados por uma aresta chamam-se **vértices vizinhos**.

A um grafo em que cada par distinto de vértices se encontra ligado por uma aresta chama-se de **grafo completo**. Num grafo completo com n vértices, o número de arestas não direcionadas é $\frac{n(n-1)}{2}$. A um grafo contido noutro chamamos de **subgrafo**.

O **peso de um vértice** de um grafo define-se pelo número de arestas que é necessário adicionar entre todos os pares de vértices seus vizinhos, de modo a que estes, o vértice e os seus vizinhos, formem um **subgrafo completo**.

Dado um grafo, pode definir-se caminhos entre vértices. Deste modo, um **caminho** $(v_1, \dots, v_n)$ define-se como a sequência de vértices distintos de um grafo, tal que $v_i \sim v_{i+1}$ para cada $i = 1, \dots, n-1$; se $v_i \rightarrow v_{i+1}$ para cada $i = 1, \dots, n-1$, ou seja, se todas as arestas forem direcionadas no mesmo sentido, diz-se esse é um **caminho direcionado** entre v_1 e v_n. Note-se que o comprimento do caminho $(v_1, \dots, v_n)$, correspondendo ao número de arestas contempladas no caminho, é $n-1$. Se for possível definir um caminho entre qualquer par de vértices que compõem um grafo, dizemos que o grafo é **conexo**. Um **ciclo** de um grafo define-se como um caminho $(v_1, \dots, v_n)$ de comprimento superior a dois onde $v_1 = v_n$. Por outras palavras, um ciclo corresponde a um caminho cuja sequência inicia e termina no mesmo vértice.

As redes bayesianas são representadas a partir de **grafos acíclicos direcionados** ou **GAD**, que são grafos com arestas direcionadas e sem ciclos direcionados.

Finalmente, e apesar desta definição não ser necessária nesta fase inicial, importa definir o conceito de **grafo em árvore** que corresponde a um grafo conexo $G = (V, A)$ não direcionado, onde para qualquer par de vértices é possível definir um único caminho entre estes. Nestes grafos as arestas são também denominadas de **ramos** e os vértices sem sucessores são chamados de **vértices folha**.

Vértices

Representadas por GAD, as estruturas gráficas das redes bayesianas são constituídas por vértices que, como já foi referido, representam as variáveis aleatórias**,** correspondentes aos conceitos do domínio do conhecimento modelado. Como já referido anteriormente, nesta dissertação estas são variáveis discretas, tomando um número finito de estados (níveis, valores, escolhas ou opções), que são mutuamente exclusivos e constituem os seus suportes. Tal não impede a consideração de problemas que envolvam variáveis aleatórias contínuas, mas significa que, caso tal aconteça, será necessário categorizar tais variáveis num número finito de categorias. Note-se que se pode saber ou não o estado em que se encontra cada

variável. O conhecimento do estado duma variável, pela sua observação, constitui a **evidência**. Esta influencia o processo de inferência. Estes conceitos serão retomados na secção 3.3.

Existem também redes bayesianas que consideram variáveis com suportes de estados contínuos, contudo essas não serão consideradas nesta dissertação. Estas são apresentadas e utilizadas, por exemplo, por Kjærulff e Madsen (2008), que apresentam e desenvolvem o conceito de redes bayesianas gaussianas.

Arestas

Koller e Friedman (2009) e Scutari e Denis (2014) referem ao longo das suas obras que, formalmente, as arestas direcionadas que ligam os vértices dos GAD configuram na maioria dos casos relações de dependência direta entre os mesmos.

Assim, uma ligação $v_1 \rightarrow v_2$ expressa uma relação de dependência direta da variável v_2 relativamente à variável v_1. Deste modo, um grafo bayesiano é composto por várias ligações direcionadas entre vértices, descrevendo interações locais entre as variáveis, da seguinte forma ilustrada:

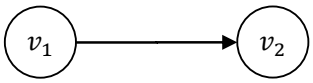


Figura 2.1 – Exemplo de uma relação de dependência definida por uma aresta direcionada.

Dado um par de vértices (v_1, v_2) de uma rede (*cf.*, *e.g.*, a Figura 2.1), dizemos que v_1 é o **antecessor** de v_2, e v_2 é o **sucessor** de v_1. Designamos por **vértices ascendentes** de um vértice v todos os vértices que são antecessores deste ou são antecessores de vértices antecessores. Os vértices sucessores e seus sucessores são chamados de **vértices descendentes**.

Causalidade e importância dos grafos na representação da causalidade

Scutari e Denis (2014) argumentam que as relações expressas pelas arestas são de mera dependência, justificando que a existência de classes de equivalências de redes, que não são distinguíveis de um ponto de vista probabilístico, prova que as direções das arestas não são indicativas de relações de causa e efeito.

Por outro lado, Kjærulff e Madsen (2008) defendem outra visão, assumindo de certa forma o conceito de causalidade na construção destas redes, uma vez que interpretam as relações expressas pelas arestas como relações de causa e efeito difusas (*fuzzy*).

Já Koller e Friedman (2009) assumem ambas as visões. Por um lado, estes autores focam as suas definições formais em propriedades probabilísticas, referindo que as direções das arestas não têm de ter propriamente um significado, uma vez que a partir de qualquer ordem das variáveis se pode criar uma rede bayesiana para qualquer distribuição. Por outro lado, tanto Koller e Friedman (2009) como Scutari e Denis (2014) referem a visão de Pearl (2009) defende de que, de um ponto de vista intuitivo, uma "boa" rede bayesiana expressa relações de causa e efeito entre as variáveis. Os autores salientam ainda o facto dos modelos construídos com base nesse pressuposto corresponderem a grafos mais esparsos e naturais, resultando numa interpretação mais clara e significativa.

Modelos bayesianos em que as arestas têm significado causal são chamados **modelos causais**, sendo os restantes chamados **modelos probabilísticos.**

Pearl (2009), como já acima referido, defende a filosofia de que as redes bayesianas expressam o funcionamento do mundo através de relações de causalidade, que têm como resultado julgamentos de independência condicional. Williamson (2002) refere que é assumido de um modo geral que grafos causais têm a simplicidade suficiente que permite reduções satisfatórias da complexidade. Segundo Pearl, a importância deste conceito advém do facto deste fornecer um formalismo de fácil compreensão e aplicação, no que toca à construção manual de modelos do conhecimento, baseando-se no padrão intuitivo de raciocínio humano em geral, que funciona através do estabelecimento de relações de causa e efeito entre variáveis. Esta forma de representação do conhecimento permite facilitar a estruturação do raciocínio de especialistas de áreas particulares do conhecimento por pessoas que não dominam essas mesmas áreas, através da especificação a partir de regras de causa e efeito. Esta dinâmica permite a redução do número de variáveis necessárias, possibilitando a construção de modelos mais compactos. A causalidade facilita não só a construção dos modelos, como também a interpretação do que estes codificam por parte de quem não acompanhou o referido processo de construção. Deste modo, e apesar de não ser aqui tratado como um conceito matemático, a representação de relações causais facilita o processo de estimação da distribuição conjunta das variáveis da rede. Note-se, contudo, que a construção de modelos causais apenas a partir de bases de dados é uma tarefa desafiante. Mais alguma informação sobre este tema pode ser encontrada no capítulo 4.

Relações de dependência e de independência

Passemos à apresentação da notação das relações expressas pelas arestas, entre as variáveis associadas a cada vértice, interpretadas como **relações de dependência** direta. Dadas duas variáveis X e Y, se estas se encontram ligadas por uma aresta, então há entre elas uma relação de dependência. De acordo com a notação apresentada na secção 2.2, referente aos conceitos da área da probabilidade, uma relação de independência entre duas variáveis denota-se pelo símbolo $\perp$. Neste sentido, e perante o cenário apresentado, escrevemos $X \not\perp Y$. Estas relações são cruciais na ótica bayesiana, no sentido em que a inferência consiste no cálculo da probabilidade de variáveis não observadas a partir de outras cujo estado foi observado.

Acresce ainda, segundo Koller e Friedman (2009), apesar da independência ser uma propriedade útil, não ser comum encontrar eventos independentes no sentido em que a situação mais comum é a verificação da independência de dois eventos dado um terceiro. Deste modo, para além das relações de dependência, identificadas entre cada par de vértices ligados por uma aresta, a partir de um GAD podem-se identificar relações de **independência condicional**. Estas relações são estabelecidas entre variáveis cujos vértices não estão ligados diretamente por uma aresta. Assim, dadas duas variáveis, X e Y, não ligadas diretamente, entre estas pode haver uma relação de independência condicionada a uma terceira variável Z, que se encontra no caminho não direcionado entre as mesmas, denotando-se esta relação por $X \perp Y|Z$.

Fluxo de informação em caminhos triplos

Consoante apresentado, a definição de uma relação de independência condicional exige a especificação de, no mínimo, três variáveis. A introdução de evidência relativa a uma das variáveis incluídas no caminho que liga o par de variáveis, das quais se pretende averiguar a relação de independência condicional, pode bloquear a transmissão de influência entre esse mesmo par. Assim, o processo de verificação de relações de independência condicional, coincide com a compreensão do fluxo de informação entre as mesmas. Entre cada conjunto de três variáveis, nomeadamente X, Y e Z, existem três tipos de conexão possível que são **ligações em série**, conexões **divergentes** e **convergentes**. Antes de apresentar estes três tipos de conexão, note-se que dadas as direções das arestas, as redes permitem três formas de raciocínio: a dedutiva, a abdutiva e a diagnóstica. A **dedução dedutiva**, ou causal, segue a direção das ligações causais entre as variáveis, e as **abdutiva** e **diagnóstica** seguem a direção oposta. A dedução diagnóstica, referida frequentemente por **inferência intercausal** refere-se à

diminuição na crença das hipóteses concorrentes uma vez observada a ocorrência de uma hipótese ou várias. A presença deste tipo de deduções nos três tipos de conexão acima referidos é apresentada nos seguintes parágrafos. Estas noções serão aprofundadas na subsecção 3.1.1.

A Figura 2.2 representa uma **ligação em série** das variáveis X, Y e Z. Nesta, a variável X tem influência sobre a variável Y e esta sobre a variável Z. Uma dada informação sobre X vai influenciar a distribuição de probabilidade da variável Y e da de Z (dedução dedutiva) e informações acerca de Z também influenciam as distribuições de probabilidade de Y e de X (dedução abdutiva). Dizemos que há então um **caminho ativo** entre X e Z. Contudo, se conhecemos o estado de Y, então a ligação entre X e Z é interrompida, e estas variáveis tornam-se independentes, e o **caminho** entre elas diz-se **inativo**. Nesse caso, temos que X e Z são independentes dado Y e escrevemos $X \perp Z|Y$.

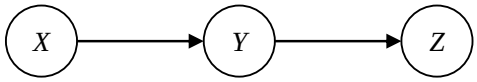


Figura 2.2 – Conexão em série entre as variáveis X, Y e Z.

Na **conexão divergente** representada na Figura 2.3, informações acerca do estado da variável X influenciam a distribuição de probabilidade da variável Z, e vice-versa. Note-se que em ambos os casos, entre X e Y e entre Z e Y se verificam os tipos de dedução dedutiva e abdutiva. O caminho entre as variáveis X e Z é classificado como ativo nesse caso. Isso só não acontece caso se conheça o estado de Y. Dessa forma, o fluxo de informação entre os descendentes é interrompido dado o seu único antecessor neste tipo de ligações, tornando-se um caminho inativo. Temos então que, tal como no caso anterior, as variáveis X e Z são independentes dada a observação do estado de Y e podemos escrever $X \perp Z|Y$.

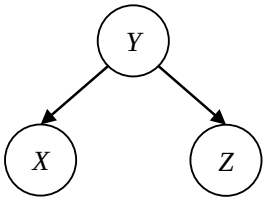


Figura 2.3 – Conexão divergente entre as variáveis X, Y e Z.

Nas **conexões convergentes**, também conhecidas por **estruturas em V**, como a da Figura 2.4, o raciocínio é oposto. Se nada se sabe acerca do estado variável intermédia Y, então os seus antecessores são independentes. Nesse caso, o caminho entre X e Z é classificado como inativo. Assim, informações acerca do estado de X não têm influência na

distribuição de probabilidade de Z. Caso se souber algo sobre Y, ou sobre um dos seus sucessores, neste caso Y_1, então X e Z tornam-se dependentes e o caminho ativo. Este é o tipo de conexão associado à inferência intercausal, abordada com maior detalhe na subseção 3.1.1. Deste modo, a única forma de ativar a estrutura em V, ou seja, de transmitir informação entre os antecessores da variável de conexão, é observar o estado dessa ou o de uma das suas sucessoras. Temos então sem a observação de Y, e das suas variáveis sucessoras, que $X \perp Z$.

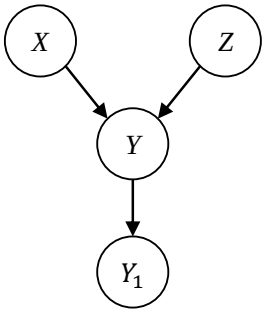


Figura 2.4 – Conexão convergente entre as variáveis X, Y e Z, tendo Y uma variável sucessora Y_1 .

Passemos a ilustrar o funcionamento da transmissão de evidência entre três variáveis ligadas numa rede causal pelos três tipos de ligação apresentados. Consideremos então o seguinte modelo composto por sete variáveis, sendo essas "*Ano de Eleições Políticas*" (E), "*Obras na Estrada*" (O), "*Más Condições Climatéricas*" (M), "*Trânsito Congestionado*" (T), "*Acidente Rodoviário*" (A), "*Polícia na Estrada*" (P) e "*Sinalização de Congestionamento*" (S).

Suponhamos que estas têm apenas dois estados, sendo esses *Sim* (s) e *Não* (n) e que o fluxo de probabilidades entre estas variáveis é representado pelo grafo presente na Figura 2.5.

Suponhamos que uma pessoa está em casa e não observou nenhum dos acontecimentos. Se for informada de que há obras na estrada (*i.e.*, que $O = s$), então é maior a sua crença de que o trânsito está congestionado ($T = s$), e deste modo, aumentará também a sua crença relativamente à presença de sinalização de congestionamento do trânsito ($S = s$). Por outro lado, se a pessoa souber da presença de sinalização de congestionamento do trânsito irá deduzir que provavelmente o trânsito está de facto congestionado ($T = s$), o que aumenta a sua crença de haver obras na estrada ($O = s$).

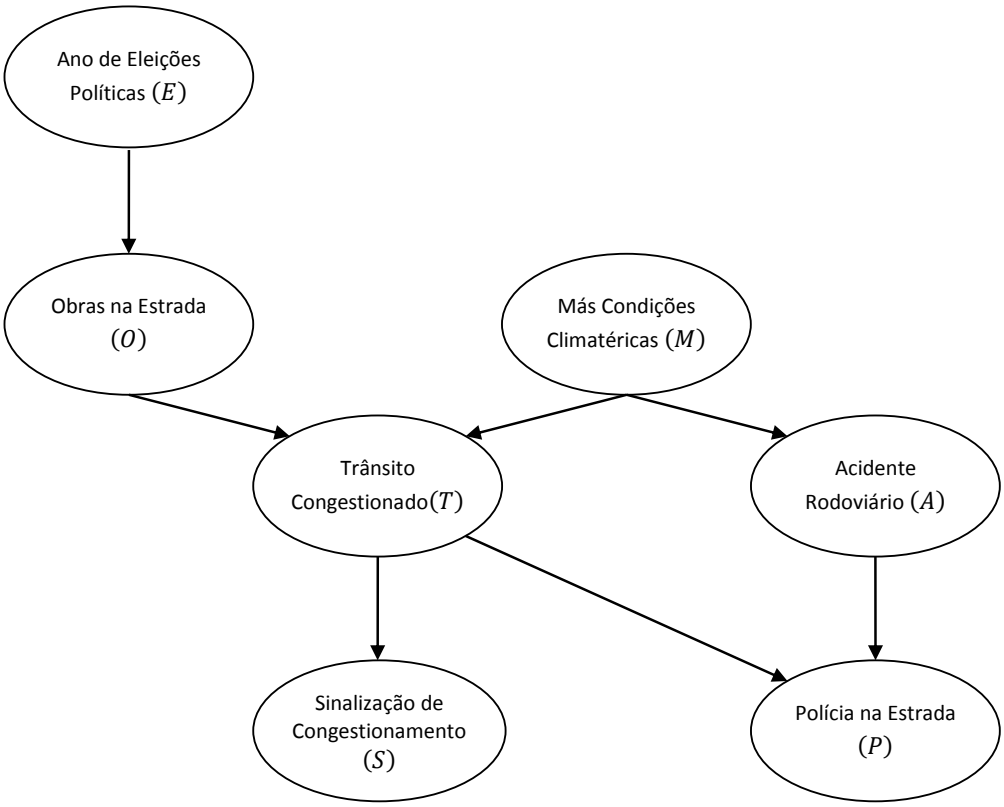


Figura 2.5 – Exemplo de rede bayesiana que modela o estado do trânsito.

Contudo, se ouvir na rádio que o trânsito está de facto congestionado, as varáveis O e S passam a ser independentes. Neste caso, se a pessoa tiver informação que de facto há sinalização de trânsito congestionado, nada pode inferir acerca de haver obras na estrada, tal como se souber que há obras na estrada, a sua crença acerca de haver sinalização de congestionamento do trânsito não é influenciada. O caminho entre O e S é então classificado como inativo se for observado o estado da variável T. Então as variáveis O e S são independentes, dada a variável T, *i.e.*, $O \perp S|T$. Este caso exemplifica o raciocínio da transmissão de informação entre os vértices na conexão em série representada pela Figura 2.6.

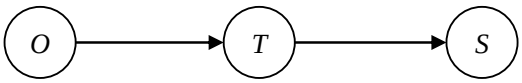


Figura 2.6 – Conexão em série entre as variáves O, T e S da rede bayesiana que modela o estado do trânsito.

Para exemplificar o caso de uma conexão divergente podemos avaliar o raciocínio entre as variáveis M, T e A, conforme a Figura . Assim, se a pessoa ouviu um vizinho dizer que o trânsito estava congestionado ($T = s$), então não sabendo nada acerca das condições

climatéricas, a sua crença de estas estarem más também aumenta ($M = s$), o que por sua vez aumenta a crença de haver um acidente ($A = s$). Por outro lado, se sabe que houve um acidente ($A = s$), a sua crença acerca das más condições temporais aumenta ($M = s$), aumentando por sua vez a crença acerca do congestionamento do trânsito ($T = s$). Mas, uma vez sabendo as condições climatéricas, as ligações entre T e A são bloqueadas, passando estas variáveis a serem independentes. Desta forma, informação acerca de A não vai influenciar a crença acerca de T, e o contrário também acontece. Dizemos então que as variáveis A e T são independentes, dada a observação do estado da variável M, *i.e*,. $A \perp T|M$.

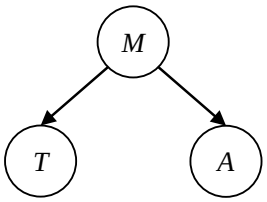


Figura 2.7 – Conexão divergente entre as variáves T, M e A da rede bayesiana que modela o estado do trânsito.

Por último, observamos a conexão existente entre os vértices O, T e M, conforme a Figura . Esta é conhecida como **conexão convergente ou estrutura V**. Neste caso, se a pessoa souber que há obras na estrada, mas não souber do estado do trânsito, a informação acerca de O não influencia a sua crença sobre o estado das condições climatéricas. O contrário também se verifica, e assim as variáveis O e M são independentes. Estas só se tornam dependentes, ou seja, só flui informação duma para outra, através de T se este vértice, ou um dos seus sucessores, neste caso S ou P, forem observados. Imaginemos que a pessoa soube que a sinalização de trânsito congestionado está ligada; então deduz que o trânsito está de facto congestionado ou que há polícia na estrada, e se souber que há obras na estrada a sua crença acerca das más condições climatéricas vai diminuir, pois vai deduzir que o aumento do trânsito se deve à ocorrência de obras. Assim conclui-se que as variáveis O e M são independentes se não se souber o estado da variável T ou de variáveis suas sucessoras.

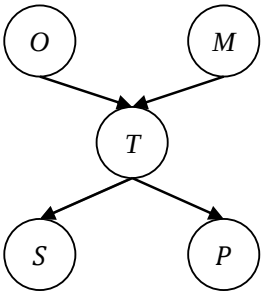


Figura 2.8 – Conexão convergente entre as variáves O, T e M da rede bayesiana que modela o estado do trânsito, com as variáveis sucessoras da variável T.

Identificação de caminhos ativos

Até aqui foram apenas considerados exemplos de **caminhos triplos**, ou seja, caminhos compostos por apenas três variáveis. Contudo, a avaliação da relação de independência condicional pode ser estendida a qualquer par de variáveis da rede. Consideremos duas variáveis X_i e X_j e a evidência formada pelo conjunto de variáveis observadas $\left(X_{k_1}, \dots, X_{k_n}\right)$. Este caso exige a consideração de todos os caminhos (não direcionados) que ligam X_i e X_j. Uma vez identificados esses caminhos, que poderão ter dimensão superior a três, deve-se passar à divisão desses, em todos os possíveis caminhos triplos. Depois, cada triplo deve ser classificado como ativo ou inativo, dada a evidência $\left(X_{k_1}, \dots, X_{k_n}\right)$. Caminhos ativos entre X_i e X_j são então compostos apenas por triplos ativos. Para que um caminho seja considerado inativo, basta que um dos caminhos triplos que o compõem seja classificado como inativo. Se todos os caminhos identificados entre X_i e X_j são classificados como inativos, dada a evidência $\left(X_{k_1}, \dots, X_{k_n}\right)$, dizemos que X_i e X_j são **separados-*d***, dada a referida evidência, estando neste caso garantida a independência condicional, ou seja, $X_i \perp X_j | \left(X_{k_1}, \dots, X_{k_n}\right)$. Caso contrário, a identificação de um ou mais caminhos ativos entre essas variáveis não garante a independência entre elas, podendo estas ser ou não independentes quando condicionadas a $\left(X_{k_1}, \dots, X_{k_n}\right)$.

Exemplo de separação-d

Analisemos, a título de exemplo, na rede bayesiana representada pela Figura 2.5 outras relações de independência condicional de variáveis, considerando variáveis que são ligadas por caminhos com dimensões superiores a três.

Suponhamos que queremos analisar se $O \perp P | A$, ou seja, se variáveis O e P são separadas-*d* dada a observação do estado da variável A. Deve-se começar por identificar todos os caminhos não direcionados que ligam O a P, que neste caso são (O, T, P) e (O, T, M, A, P). Depois identificam-se todos os triplos em que se podem dividir cada um destes caminhos. O primeiro já constitui um caminho triplo, $(O \longrightarrow T \longrightarrow P)$, que é classificado como ativo por ser uma conexão em série em que a variável intermédia, *i.e.*, T, não foi observada. Deste modo, o caminho (O, T, P) é classificado como ativo, sendo tal suficiente para dizer que O e P não são separados-*d* dado A. No entanto vamos continuar a análise do segundo caminho, (O, T, M, A, P), que pode ser decomposto nos triplos $(O \longrightarrow T \longleftarrow M)$, $(T \longleftarrow M \longrightarrow A)$ e $(M \longrightarrow A \longrightarrow P)$. O primeiro e o último são classificados como inativos por corresponderem, respetivamente, a uma conexão convergente onde a variável intermédia, *i.e.*, T, não foi

observada, e uma conexão em série em que variável intermédia, *i.e.*, A, foi observada. Assim, embora o segundo triplo seja ativo, pois a variável M não foi observada, este segundo é um caminho classificado como inativo.

2.4 Fatorização das redes bayesianas

Uma consequência importante da separação-*d* é o facto de, para qualquer $i \in \{1, \dots, n\}$, a variável X_i, dadas as suas antecessoras ser independente das suas variáveis não descendentes. Note-se que se exclui a variável X_i das suas variáveis não descendentes.

Deste modo, o grafo G codifica as seguintes relações chamadas de **independências locais**:

$$X_i \perp Não\ descendentes\ de\ X_i \mid \text{ant}(X_i), \tag{2.38}$$

onde $\text{ant}(X_i)$ denota o conjunto das variáveis antecessoras de X_i.

Exemplo

Verifiquemos as consequências da aplicação do resultado anterior à rede representada pela Figura 2.5. Por exemplo, no caso da variável S, a observação do estado da sua única antecessora T, implica a sua independência condicional das restantes variáveis da rede. Usando a notação $X \perp \{Y_1, \dots, Y_n\} \mid Z$ para representar $X \perp Y_i \mid Z, \forall\, i \in \{1, \dots, n\}$, temos então:

$$S \perp Não\ descendentes\ de\ S \mid \text{ant}(S) \Leftrightarrow S \perp \{E, O, T, M, A, P\} \mid T. \tag{2.39}$$

Assim, a expressão (2.38) que estabelece a independência condicional entre cada variável e as suas não descendentes tem implicações em termos da distribuição probabilística conjunta. A sua utilidade advém da redução das probabilidades condicionais de cada variável às das suas antecessoras. De facto, denotando por $X_1, \dots, X_{i-1}$ as variáveis não descendentes de X_i,

$$P(X_i \mid X_1, \dots, X_{i-1}) = P\big(X_i \mid \text{ant}(X_i)\big). \tag{2.40}$$

No exemplo da Figura 2.5 temos que $P(S \mid E, O, M, A, P, T) = P(S \mid T)$.

Como consequência, e atendendo à regra da cadeia, a distribuição conjunta das variáveis que compõem a rede pode ser reduzida a:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P\big(X_i \mid \text{ant}(X_i)\big), \tag{2.41}$$

onde, no caso de X_i não ter antecessores, se convenciona que $P\big(X_i|ant(X_i)\big) = P(X_i)$.

Assim, a distribuição conjunta da rede bayesiana apresentada como exemplo da Figura 2.5, pode ser fatorizada da seguinte forma:

$$P(E,O,M,T,A,S,P) = P(E)P(O|E)P(M)P(T|O,M)P(S|T)P(A|M)P(P|A,T). \quad (2.42)$$

Uma vez que o modelo é composto por 7 variáveis, em vez de especificar a probabilidade conjunta $P(E,O,M,T,A,S,P)$ por uma tabela de $2^7 = 128$ entradas, as relações de causa estabelecidas pela estrutura gráfica e as subsequentes relações de independência condicionais implicadas, permitem a redução dessa distribuição a 7 outras de menor dimensão. São essas as distribuições de cada variável condicionada às suas antecessoras, nomeadamente $P(E)$, $P(O|E)$, $P(M)$, $P(T|O,M)$, $P(S|T)P(A|M)$, e $P(P|A,T)$.

É deste modo que o conceito não matemático de causalidade, subjacente às ligações estabelecidas entre as variáveis, ajuda na simplificação e compactação da especificação da probabilidade conjunta de variáveis. Mediante a verificação das relações de independência condicional que os modelos gráficos das redes bayesianas expressam, pela junção das relações causais locais, o número de fatorizações associadas a um conjunto de variáveis pode ser reduzido àquelas que traduzem as relações estabelecidas no grafo. No capítulo 4 é abordada a verificação das independências condicionais na prática.

Resumindo, uma rede bayesiana pode definir-se por $B = (\boldsymbol{X}, G, \boldsymbol{P})$, constituída por um conjunto de variáveis aleatórias $\boldsymbol{X} = (X_1, \dots, X_n)$, um grafo acíclico direcionado $G = (V, A)$, que as representa pelos seus vértices V, e as liga por arestas direcionadas A, estando a cada variável X_i associada uma distribuição probabilística condicional $P \subseteq \boldsymbol{P}$, que se pode escrever na forma $P\big(X_i|\mathrm{ant}(X_i)\big)$. Então a semântica das redes bayesianas resume-se aos seguintes três pontos:

1. Um GAD representado por G, com um vértice por variável X_i;
2. Uma tabela de distribuição de cada variável, condicionada às suas antecessoras: $P\big(X_i|\mathrm{ant}(X_i)\big)$;
3. A distribuição conjunta das variáveis, codificada pelo produto das distribuições associadas a cada uma: $P(X_1, \dots, X_n) = \prod_{i=1}^{n} P\big(X_i|\mathrm{ant}(X_i)\big)$.

Exemplo

Consideremos novamente a rede representada pela Figura 2.5. Tem-se $B = (\boldsymbol{X}, G, \boldsymbol{P})$, estando a cada variável contida em $\boldsymbol{X}$ e representada em G, ou seja, a E, O, M, T, A, S e P,

associada uma distribuição probabilística condicional às variáveis antecessoras, nomeadamente $P(E)$ $P(O|E)$, $P(M)$, $P(T|O,M)$, $P(A|M)$, $P(S|T)$ e $P(P|A,T)$, distribuições que são especificadas nas tabelas representadas na Figura 2.9,

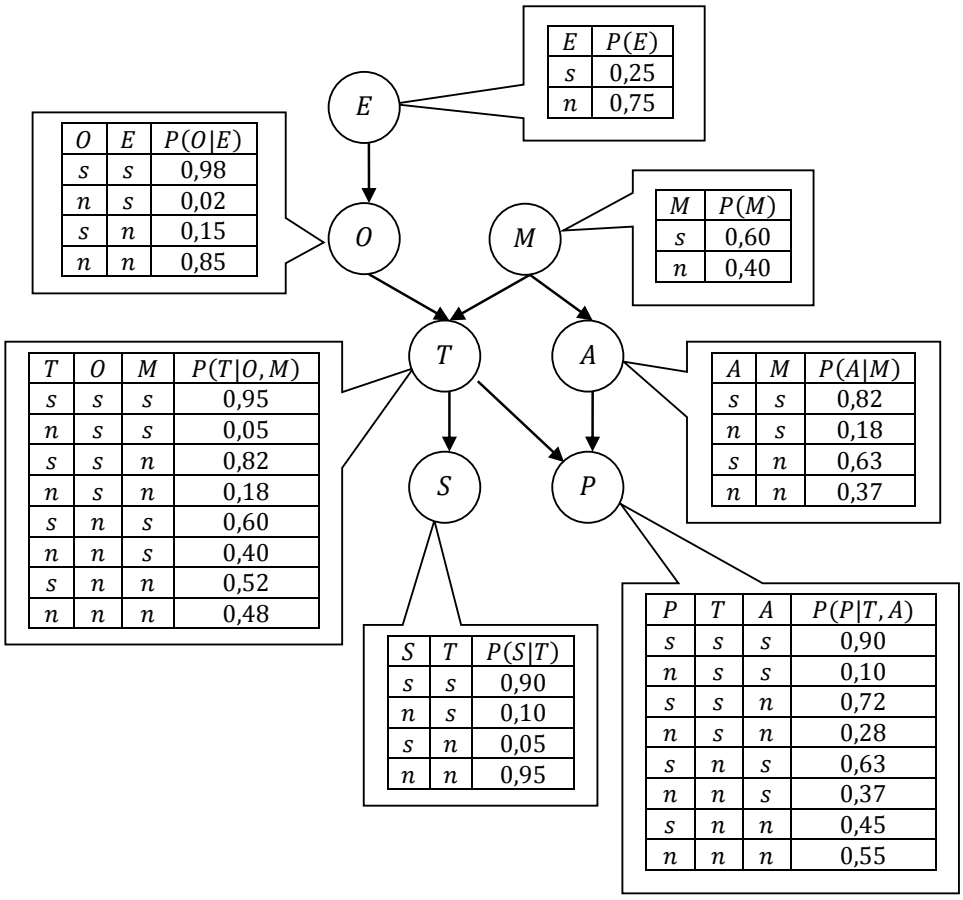


Figura 2.9 – Exemplo da rede bayesiana do estado do trânsito, com as tabelas de probabilidade condicionais associadas.

Deste modo, se observarmos o estado de cada variável, por exemplo $E = s$, $O = n$, $M = s$, $T = s$, $O = s$, $A = n$, $S = n$ e $P = s$, baseados nas tabelas de distribuição condicional apresentadas podemos calcular a probabilidade conjunta de observação desses estados da forma apresentada em (2.43).

$$\begin{aligned} P(E=s,O=n,M=s,T=s,A=n,S=n,P=s) &= P(E=s)P(O=n|E=s) \\ &\quad P(M=s)P(T=s|O=n,M=s) \\ &\quad P(A=n|M=s)P(S=n|T=s) \\ &\quad P(P=s|A=n,T=s) \\ &= 0{,}25\times0{,}02\times0{,}6\times0{,}6\times0{,}18\times0{,}1\times0{,}72 \\ &\approx 0.0000233. \end{aligned} \tag{2.43}$$

2.5 Exemplos de redes

Posteriormente ao desenvolvimento das redes bayesianas na década de 1980, nessa mesma década, e nas décadas seguintes, foram publicados inúmeros estudos de áreas diferentes propondo as redes bayesianas como instrumento de modelação. As aplicações práticas dos modelos existentes vão desde a medicina à biologia, passando pela agricultura, hidrografia, meteorologia e economia. Alguns modelos são de seguida apresentados.

Na área da medicina estas redes apresentam-se como ótimos instrumentos de diagnóstico de doenças possíveis dados os sintomas observados, ou de auxílio na escolha dos melhores exames a realizar dados os sintomas e possíveis doenças associadas aos mesmos. Nesta área salientam-se os sistemas de grande dimensão *Pathfinder* com 109 vértices e 195 arestas e *Munin* com 1041 vértices e 1397 arestas.

A rede *Pathfinder* foi apresentada por Heckerman *et al.* (1992), tendo sido desenvolvida para apoiar patologistas cirúrgicos no diagnóstico de doenças na glândula linfática. No artigo referido, os autores mostram que o modelo é tratável e expressivo o suficiente para capturar as características fundamentais da patologia das glândulas linfáticas.

Por outro lado, a rede *Munin* é um modelo que foi construído com o intuito de diagnosticar distúrbios neuromusculares a partir de dados adquiridos por eletromiografia e apresentado no endereço repositório de redes bayesianas dos autores de Scutari e Denis (2014)[3]. Segundo Andreassen *et al.* (1996), a capacidade do diagnóstico da rede Munin está ao mesmo nível de um especialista.

Na área da biologia encontramos o modelo *Sachs*, que é de pequena dimensão com 11 vértices e 17 arestas, e que foi construído com o objetivo de ser capaz de representar relações diretas e indiretas complexas entre várias moléculas. Esta sistema funciona através da medida simultânea de 11 proteínas fosforiladas e componentes fosfolipídicos, em milhares de células humanas primárias do sistema imunitário sujeitas a intervenções moleculares gerais e específicas. Este modelo foi apresentado por Sachs *et al.* (2005).

Na área da agricultura destacamos o modelo *Barley*, com 48 vértices e 84 arestas, que se insere numa vertente de produção agrícola biológica, tendo sido desenvolvido no âmbito do projeto "Produção de cerveja a partir de cevada maltada cultivada sem pesticidas" de Kristensen e Rasmussen (2002).

[3] http://www.bnlearn.com/bnrepository/ (consultado em 3 de maio de 2019).

O repositório de redes bayesianas refere ainda o modelo *Water*, com 39 vértices e 66 arestas, que se insere na área da hidrografia, tendo sido apresentado num relatório intitulado *An Expert System for Control of Waste Water Treatment - A Pilot* em 1989. Este refere-se a um projeto piloto de sistemas especializados desenhados para o controlo do desperdício de água.

O modelo *Hailfinder*, com 56 vértices e 66 arestas, insere-se na área da meteorologia, e foi desenhado para a previsão climatérica severa no norte do Estado do Colorado, tendo sido apresentado por Abramson *et al*. (1996).

Finalmente na área económica, mais especificamente dos seguros, encontramos o modelo *Insurance*, com 27 vértices e 52 arestas, que auxilia as seguradoras automóveis a avaliar o risco de um potencial cliente, dadas várias variáveis como a idade do mesmo, as suas capacidades de condução, o seu historial de acidentes e várias características do carro. Este foi apresentado em Binder *et al.* (1997), onde são apresentados algoritmos de estimação de redes com variáveis não observadas.

3 Inferência

Scutari e Denis (2014) referem que as redes bayesianas são usadas para responder a questões acerca da natureza dos seus dados, as chamadas **questões parciais** (*queries*). Essas são obtidas através de técnicas conhecidas por **inferência**, **raciocínio probabilístico** ou **atualização de crenças**. Os autores referem que estes termos originais dos sistemas especializados foram introduzidos na terminologia probabilística por Pearl (1988), ressaltando o seu uso em obras recentes como a de Koller e Friedman (2009).

Dada uma rede bayesiana $B = (\boldsymbol{X}, G, \boldsymbol{P})$ com n variáveis, $(X_1, \dots, X_n)$, a resolução de uma questão parcial $Q = (B, \boldsymbol{A}, \boldsymbol{E})$ consiste no cálculo da probabilidade condicional $P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e})$ de um conjunto de **variáveis alvo** cujo o estado não foi observado, denominado por $\boldsymbol{A}$, dada a evidência $\boldsymbol{E}$, formada pelo conjunto das k variáveis cujo estado foi observado $\boldsymbol{E} = \boldsymbol{e}$, *i.e.*, $\boldsymbol{E} = (E_1 = e_1, \dots, E_k = e_k)$. As variáveis de $\boldsymbol{X}$ não incluídas em $\boldsymbol{A}$ nem em $\boldsymbol{E}$, chamam-se ocultas e formam o conjunto $\boldsymbol{O}$.

3.1 Tipos de questões parciais

Scutari e Denis (2014) e Koller e Friedman (2009) identificam duas categorias principais de questões parciais, nas quais focam a sua obra. São essas as **questões parciais de probabilidade condicional** (CPQ, do inglês, *Conditional Probability Query*), **questões parciais máximas *a posteriori*** (MAP, do inglês, *Maximum a Posteriori*), e **questões parciais de explicação mais provável** (MPE, do inglês, *Most Probable Explanation*), ou ***MAP marginais.***

3.1.1 Questões parciais de probabilidade condicional (CPQ)

Consoante o tipo de dedução realizada Koller e Friedman (2009) classificam as **questões parciais** de probabilidade condicional em três categorias: de **raciocínio causal** ou **de previsão**, **de raciocínio de evidência** ou **de explicação** e de **raciocínio intercausal**.

Questões parciais de **raciocínio causal** ou **de previsão** preveem o efeito a jusante de vários fatores, estimando a probabilidade de uma variável tendo em conta a observação do estado de variáveis não descendentes. Na rede bayesiana representada pela Figura 2.5, um exemplo deste tipo de questão seria a estimação do estado da sinalização de congestionamento (S), dada observação do estado do trânsito, por exemplo, $T = s$, e da variável *acidente rodoviário*, por exemplo, $A = s$. Esta questão parcial consiste então na estimação da probabilidade $P(S|T = s, A = s)$.

Nas **questões parciais** de **raciocínio de evidência** ou **de explicação** o raciocínio é feito a partir das variáveis efeito para as variáveis causa. Assim, tomando novamente como exemplo a rede representada na Figura 2.5, um exemplo deste tipo de questão poderia ser a estimação da probabilidade de haver obras na estrada (O), dado que se observou que a sinalização de congestionamento está ligada, $S = s$, e que há polícia na estrada, $P = s$, *i.e.*, $P\,(O|S = s, P = s)$.

Finalmente, nas **questões parciais** de **raciocínio intercausal,** tal como referido nos tipos de dedução, é abordada a interação de variáveis causa de uma mesma variável efeito. Considerando a rede representada na Figura 2.5, um exemplo deste tipo de questão poderia ser a estimação da probabilidade da variável *más condições climatéricas* (M), dada a observação do estado da variável efeito *trânsito congestionado* (T), e a variável causa em comum *obras na estrada*, (O), por exemplo $P\,(M|O = n, T = s)$.

3.1.2 Questões MAP e MPE

Consideremos um subconjunto de variáveis $\boldsymbol{A}$ não contidas no conjunto de evidência, ou seja, $\boldsymbol{A} \subseteq \boldsymbol{X} - \boldsymbol{E}$. Segundo Koller e Friedman (2009) as questões parciais MAP e MPE são equivalentes e a sua resolução consiste na estimação da instanciação do conjunto das variáveis $\boldsymbol{A}$, que maximiza a sua distribuição conjunta. Darwiche (2008) define separadamente estas questões, consoante o conjunto $\boldsymbol{A}$ coincide com o conjunto $\boldsymbol{X} - \boldsymbol{E}$ (MAP), ou $\boldsymbol{A}$ é um subconjunto estrito de $\boldsymbol{X} - \boldsymbol{E}$ (MPE). Assim, tal como referem Scutari e Denis (2014), a resolução deste tipo de questões parciais é encontrar uma configuração $\boldsymbol{a}^*$, das variáveis em $\boldsymbol{A}$, que maximiza a seguinte probabilidade condicional:

$$MAP(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e}) = \boldsymbol{a}^* = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, P(\boldsymbol{A} = \boldsymbol{a}|\boldsymbol{E} = \boldsymbol{e}). \tag{3.1}$$

Note-se que o resultado pode não ser único.

Existem ainda as questões parciais $\boldsymbol{MPE}$ ou $\boldsymbol{MAP}$ ***marginais*** que combinam elementos das questões parciais MAP e das CPQ. Nestas, existe um conjunto de variáveis $\boldsymbol{A}$ que formam a questão parcial. Sendo $\boldsymbol{O} = \boldsymbol{X} - \boldsymbol{A} - \boldsymbol{E}$, a tarefa destas questões parciais consiste em calcular a distribuição mais provável para as variáveis $\boldsymbol{A}$ dada a evidência $\boldsymbol{E} = \boldsymbol{e}$:

$$MPE(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e}) = \underset{\boldsymbol{A}}{\operatorname{argmax}} \sum_{\boldsymbol{O}} P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e}). \tag{3.2}$$

3.2 Resolução de CPQ

3.2.1 Inferência por enumeração

Definida uma questão parcial CPQ a resolver, temos a repartição das variáveis da rede em três grupos. São esses os das **variáveis alvo**, $\boldsymbol{A} = (A_1, \dots, A_m)$, das **variáveis evidência** $\boldsymbol{E} = (E_1, \dots, E_k)$, e das **variáveis ocultas** $\boldsymbol{O} = (O_1, \dots, O_r)$. O grupo das variáveis ocultas é composto por todas aquelas que não estão em nenhum dos outros grupos, ou seja, que não são contempladas na questão parcial. Note-se que os conjuntos de variáveis $\boldsymbol{A}, \boldsymbol{E}$ e $\boldsymbol{O}$ formam uma partição de $\boldsymbol{X}$, *i.e.*, são conjuntos disjuntos, dois a dois, e $\boldsymbol{A} \cup \boldsymbol{E} \cup \boldsymbol{O} = \boldsymbol{X}$.

Pela definição de probabilidade condicional temos:

$$P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e}) = \frac{P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e})}{P(\boldsymbol{E} = \boldsymbol{e})}, \tag{3.3}$$

da qual resulta:

$$P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e}) \propto P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e}). \tag{3.4}$$

Temos então que o valor de uma questão parcial é proporcional à distribuição conjunta das variáveis alvo e evidência. A resolução de questões parciais segue a aplicação do princípio básico de alterar a distribuição conjunta das variáveis representadas pelos vértices, incorporando a nova informação recolhida, sendo essa incorporação obtida através do condicionamento dessa distribuição aos valores das variáveis observadas.

A operação de cálculo anterior é interpretada por alguns autores, como Jensen (1996) e Kjærulff e Madsen (2008), em termos da função verosimilhança. Essa interpretação advém da igualdade

$$P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e}) = P(\boldsymbol{E} = \boldsymbol{e}|\boldsymbol{A})P(\boldsymbol{A}), \tag{3.5}$$

em conjunto com a relação

$$P(\boldsymbol{E} = \boldsymbol{e}|\boldsymbol{A}) = L(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e}), \tag{3.6}$$

justificando-se, deste modo, a seguinte relação de proporcionalidade:

$$P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e}) \propto L(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e})P(\boldsymbol{A}). \tag{3.7}$$

Em termos práticos vamo-nos basear na expressão (3.4), e na visão de Scutari e Denis (2014), resolvendo cada questão parcial a partir da distribuição conjunta das variáveis alvo $\boldsymbol{A}$ e evidência $\boldsymbol{E} = \boldsymbol{e}$, *i.e.*, $P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e})$. Esta pode obter-se da partir da distribuição conjunta de

todas as variáveis $\boldsymbol{X}$ da rede bayesiana, *i.e.,* $P(\boldsymbol{X}) = P(\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{O})$, cuja fatorização é induzida pela estrutura e pela regra de cadeia, marginalizando de seguida as variáveis $\boldsymbol{O}$, ou seja:

$$P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e}) = \sum_{\substack{\boldsymbol{O} \\ \boldsymbol{E}=\boldsymbol{e}}} P(\boldsymbol{X}) = \sum_{\substack{\boldsymbol{O} \\ \boldsymbol{E}=\boldsymbol{e}}} P(\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{O}) = \sum_{\substack{\boldsymbol{O} \\ \boldsymbol{E}=\boldsymbol{e}}} \prod_{X_i \in \boldsymbol{X}} P\big(X_i | \text{ant}(X_i)\big). \tag{3.8}$$

Uma vez calculada a probabilidade $P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e})$, esta deve ser dividida por $P(\boldsymbol{E} = \boldsymbol{e})$.

A título de exemplo, considerando novamente a rede representada na Figura 2.5, se desejamos calcular, por exemplo, $P(O|S = s, P = s)$ seguem-se os seguintes passos:

$$\begin{aligned} & P(O|S = s, P = s) = \\ & = \frac{P(O, S = s, P = s)}{P(S = s, P = s)} \\ & = \frac{\sum_{E,M,A,T} P(E, O, M, T, A, S = s, P = s)}{\sum_{E,O,M,A,T} P(E, O, M, T, A, S = s, P = s)} \\ & = \frac{\sum_{E,M,A,T} P(E)P(O|E)P(M)P(T|O, M)P(A|M)P(S = s|T)P(P = s|A, T)}{\sum_{E,O,M,A,T} P(E)P(O|E)P(M)P(T|O, M)P(A|M)P(S = s|T)P(P = s|A, T)}. \end{aligned} \tag{3.9}$$

Depois de alguns cálculos obtemos $P(O|S = s, P = s)$ (*cf.* secção 6.9).

3.2.2 Inferência pelo algoritmo da eliminação de variáveis

Koller e Friedman (2009) referem que se pode usar um modelo gráfico para responder a qualquer questão parcial, simplesmente gerando a distribuição conjunta e eliminando exaustivamente as variáveis necessárias eliminar, como aconteceu no exemplo da secção anterior. Contudo, segundo os autores, esta abordagem não é satisfatória no sentido em que a enumeração da distribuição conjunta exige um número exponencial de cálculos, invalidando o objetivo de compactação da representação gráfica. Neste sentido, os autores apresentam um algoritmo, em alternativa à abordagem anterior de enumeração, que permite a simplificação deste processo em muitos casos, que designam por **algoritmo da eliminação de variáveis** (ou E.V.).

Koller e Friedman (2009) ressaltam que a questão do crescimento exponencial de cálculos na inferência é combatido através da estrutura da rede bayesiana, que permite que as probabilidades condicionais associadas à rede dependam apenas de um grupo restrito de variáveis, e do armazenamento por etapas dos resultados parciais, evitando assim a repetição desnecessária de cálculos. Note-se que, ainda assim, existem alguns casos extremos em que a exigência de um número exponencial de cálculos não é evitada com este algoritmo. Tal

acontece, por exemplo, em modelos em que as variáveis em causa contêm suportes com um número elevado de estados.

Assim, o algoritmo de eliminação de variáveis tem como operação chave a marginalização de variáveis, manipulando fatores. Considerando $\boldsymbol{X}$ um conjunto de variáveis e Y uma variável, tal que $Y \notin \boldsymbol{X}$, Koller e Friedman (2009) definem a operação de eliminar Y de um fator $\phi(\boldsymbol{X}, Y)$, denotada por $\sum_Y \phi$, que resulta na criação de um novo fator τ, sobre $\boldsymbol{X}$, da seguinte forma:

$$\tau(\boldsymbol{X}) = \sum_Y \phi\ (\boldsymbol{X}, Y). \tag{3.10}$$

O algoritmo parte de um conjunto inicial de fatores $\boldsymbol{\Phi}$, composto pelo conjunto de probabilidades condicionais que fatorizam a probabilidade conjunta da rede, ou seja, $\boldsymbol{\Phi} = \{\phi_{X_i}\}$ onde $\phi_{X_i} = P(X_i|\text{ant}(X_i))$ e de uma ordem segundo a qual se vão eliminar as r variáveis ocultas da questão parcial, ou seja, $O = \{O_1, \dots, O_r\}$. Seguindo essa ordem, elimina-se uma variável de cada vez, seguindo os seguintes passos:

1. Para eliminar uma variável O_i selecionam-se todos os fatores ϕ do grupo $\boldsymbol{\Phi}$ que a incluem como variável condicionada ou condicionante;
2. Define-se então um novo grupo $\boldsymbol{\Phi}'$ composto por esses fatores;
3. Retiram-se os fatores selecionados do conjunto inicial obtendo-se $\boldsymbol{\Phi}''$, *i.e.*, $\boldsymbol{\Phi}'' = \boldsymbol{\Phi} - \boldsymbol{\Phi}'$;
4. Multiplicam-se todos os fatores de $\boldsymbol{\Phi}'$, gerando outro, denominado ψ, *i.e.*, $\prod_{\phi \in \Phi'} \phi = \psi$;
5. Depois, marginaliza-se a variável O_i de ψ, gerando outro fator τ, *i.e.*, $\sum_{O_i} \psi = \tau$.
6. Este último junta-se ao conjunto de fatores principal, *i.e.*, $\boldsymbol{\Phi}''$ ∪ $\{\tau\}$, passando este conjunto a ser o novo conjunto de fatores $\boldsymbol{\Phi}$.

Os passos 1-6 repetem-se r vezes, até se eliminarem todas as r variáveis pretendidas.

Podemos, a título de exemplo, aplicar este algoritmo à rede bayesiana representada na Figura 2.5. Segundo a regra da cadeia a rede define a distribuição conjunta com os fatores identificados da seguinte forma (já apresentada em (2.42)):

$$P(E, O, M, T, A, S, P) = P(E)P(O|E)P(M)P(T|O, M)P(A|M)P(S|T)P(P|A, T). \tag{3.11}$$

No caso de querermos calcular $P(P)$ através do método da enumeração (3.8):

$$\begin{aligned} P(P) &= \sum_{E,O,M,A,T,S} P(E,O,M,T,A,S,P) \\ &= \sum_{E,O,M,A,T,S} P(E)P(O|E)P(M)P(T|O,M)P(A|M)P(S|T)P(P|A,T). \end{aligned} \tag{3.12}$$

O algoritmo de E.V. elimina as variáveis uma de cada vez, por etapas, ao contrário do da enumeração que funciona marginalizando as variáveis ocultas de uma vez. Seguindo por exemplo a ordem de eliminação E, O,M, S, A, T, consideremos:

$$P(P) = \sum_{P}\left(\sum_{T}\left(\sum_{A} P(P|A,T)\left(\sum_{S} P(S|T)\sum_{M} P(M)P(A|M)\,\Theta\right)\right)\right), \tag{3.13}$$

onde,

$$\Theta = \sum_{O} P(T|O,M)\left(\sum_{E} P(E)P(O|E)\right). \tag{3.14}$$

Traduzindo as probabilidades condicionais apresentadas na expressão (3.3) em termos de fatores temos:

$$\begin{aligned} P(E,O,M,T,A,S,P) &= P(E)P(O|E)P(M)P(T|O,M)P(A|M)P(S|T)P(P|A,T) \\ &= \phi_E(E)\phi_O(O,E)\phi_M(M)\phi_T(O,M,T)\phi_A(A,M)\phi_S(S,T)\phi_P(P,A,T). \end{aligned} \tag{3.15}$$

Utilizando os fatores identificados na expressão anterior (3.15) e a ordem de eliminação E, O, M, S, A e T, apresentada nas expressões (3.13) e (3.14), o algoritmo concretiza-se de acordo com as seguintes operações:

1. Eliminar E:

$$\psi_1(E,O) = \phi_E(E)\phi_O(O,E) = P(E)P(O|E) = P(O,E); \tag{3.16}$$

$$\tau_1(O) = \sum_{E} \psi_1(E,O) = \sum_{E} P(O,E) = P(O); \tag{3.17}$$

2. Eliminar O:

$$\psi_2(O,M,T) = \tau_1(O)\phi_T(O,M,T) = P(O)P(T|O,M) = P(O,T|M); \tag{3.18}$$

$$\tau_2(M,T) = \textstyle\sum_O \psi_2(O,M,T) = \sum_O P(O,T|M) = P(T|M); \tag{3.19}$$

3. Eliminar M:

$$\begin{aligned}\psi_3(M,T,A) &= \tau_2(M,T)\phi_M(M)\phi_A(A,M) \\ &= P(T|M)P(M)P(A|M) = P(A,T|M)P(M) = P(T,M,A);\end{aligned} \tag{3.20}$$

$$\tau_3(T,A) = \sum_M \psi_3(M,T,A) = \sum_M P(T,M,A) = P(T,A); \tag{3.21}$$

4. Eliminar S:

$$\psi_4(S,T) = \phi_S(S,T) = P(S|T); \tag{3.22}$$

$$\tau_4(T) = \sum_S \psi_4(S,T) = \sum_S P(S|T) = 1; \tag{3.23}$$

5. Eliminar A:

$$\psi_5(T,A,P) = \tau_3(T,A)\phi_P(P,A,T) = P(T,A)P(P|A,T) = P(P,T,A); \tag{3.24}$$

$$\tau_5(T,P) = \sum_A \psi_5(T,A,P) = \sum_A P(P,T,A) = P(P,T); \tag{3.25}$$

6. Eliminar T**:**

$$\psi_6(T,P) = \tau_4(T)\tau_5(T,P) = 1 \times P(P,T) = P(P,T); \tag{3.26}$$

$$\tau_6(P) = \sum_T \psi_6(T,P) = \sum_T P(P,T) = P(P). \tag{3.27}$$

Note-se também que esta eliminação das variáveis ocultas poderia ser feita segundo uma ordem diferente. As Tabela 3.1 e Tabela 3.2 explicitam diferentes ordens de eliminação das variáveis ocultas para a estimação da questão parcial $P(P)$. Enquanto que a Tabela 3.1 segue a ordem de eliminação das variáveis descrita no exemplo anterior, a Tabela 3.2 apresenta uma ordem diferente. Consoante a ordem escolhida, os fatores intermédios gerados podem ter dimensões diferentes. Deste modo, em termos de armazenamento computacional e de tarefas de cálculo, é benéfico identificar a ordem de eliminação que gera fatores de menor dimensão. Koller e Friedman (2009) referem que uma possível forma de contornar esta dificuldade pode passar pelo uso de outras estruturas, nomeadamente das árvores junção, cujo algoritmo é apresentado na subsecção seguinte, 3.2.3.

Tabela 3.1 – Aplicação do algoritmo de eliminação de variáveis para determinar $P(P)$.

Passo	*Variável Eliminada*	*Fatores usados*	*Variáveis Envolvidas*	*Fator Novo*
1	E	$\phi_E(E), \phi_O(O,E)$	E,O	$\tau_1(O)$
2	O	$\tau_1(O), \phi_T(O,M,T)$	O,M,T	$\tau_2(M,T)$
3	M	$\tau_2(M,T), \phi_M(M), \phi_A(A,M)$	M,T,A	$\tau_3(T,A)$
4	S	$\phi_S(S,T)$	S,T	$\tau_4(T)$
5	A	$\tau_3(T,A), \phi_P(P,A,T)$	T,A,P	$\tau_5(T,P)$
6	T	$\tau_4(T), \tau_5(T,P)$	T,P	$\tau_6(P)$

Tabela 3.2 – Aplicação do algoritmo de eliminação de variáveis para determinar $P(P)$, com uma ordem diferente de eliminação de variáveis.

Passo	*Variável Eliminada*	*Fatores usados*	*Variáveis Envolvidas*	*Fator Novo*
1	M	$\phi_T(O,M,T), \phi_M(M), \phi_A(A,M)$	O,M,A,T	$\tau_1(O,A,T)$
2	E	$\phi_E(E), \phi_O(O,E)$	E,O	$\tau_2(O)$
3	S	$\phi_S(S,T)$	S,T	$\tau_3(T)$
4	O	$\tau_2(O), \tau_1(O,A,T)$	O,T,A	$\tau_4(T,A)$
5	T	$\tau_4(T,A), \tau_3(T)\phi_P(P,A,T)$	T,A,P	$\tau_5(A,P)$
6	A	$\tau_5(A,P)$	A,P	$\tau_6(P)$

Introdução de evidência

Note-se que o exemplo anterior não inclui variáveis cujo estado foi observado. Deste modo Koller e Friedman (2009), referem que a introdução da evidência passa apenas pelo cálculo de $P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e})$, aplicando o algoritmo de eliminação de variáveis ao conjunto $\boldsymbol{O} = \boldsymbol{X} - \boldsymbol{A} - \boldsymbol{E}$. O fator final obtido pelo algoritmo apresentado anteriormente, depois de substituir $\boldsymbol{E}$ por $\boldsymbol{e}$, $\phi'(\boldsymbol{A})$, corresponde a $P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e})$**.** Para se obter $P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e})$ a partir desse, seguindo a expressão (3.3), basta normalizar o fator $\phi'(\boldsymbol{A})$, multiplicando-o por $\frac{1}{\alpha}$, onde α corresponde à soma das entradas da distribuição $P(\boldsymbol{A}, \boldsymbol{E} = \boldsymbol{e})$ não normalizada, que corresponde à probabilidade da evidência, *i.e.*, $P(\boldsymbol{E} = \boldsymbol{e})$**.**

A título de exemplo, suponhamos que o objetivo é calcular $P(P|O = s, A = n)$. Começamos por determinar a distribuição $P(P, O = s, A = n)$, que normalizaremos posteriormente**.** Uma forma possível de resolução deste problema é apresentada na Tabela 3.3 que segue a mesma ordem de eliminação de variáveis apresentada na Tabela 3.2, com a exclusão das variáveis O e A que foram observadas. . Note-se que τ'_2 é um número que deve ser introduzido num dos passos subsequentes para que a evidência $O = s$ esteja presente no resultado final.

Tabela 3.3 – Aplicação do algoritmo de eliminação de variáveis para a determinação da questão parcial $P(P, O{=}s, A{=}n)$.

Passo	*Variável Eliminada*	*Fatores usados*	*Variáveis Envolvidas*	*Fator Novo*
1	M	$\phi_T[O=s](M,T), \phi_M(M), \phi_A[A=n](M)$	M,T	$\tau'_1(T)$
2	E	$\phi_E(E), \phi_O[O=s](E)$	E	τ'_2
3	S	$\phi_S(S,T)$	S,T	$\tau'_3(T)$
4	T	$\tau'1(T), \tau'2, \tau'3(T), \phi P\ [A=n](P,T)$	T,P	$\tau'_4(P)$

3.2.3 Algoritmo Hugin: árvores junção

A inferência sobre redes bayesianas pode também ser efetuada de forma eficiente sobre uma estrutura secundária, conhecida pela designação **árvore de junção** (*junction tree* ou *joint tree*) ou **árvore clique.**

Fenton e Neil (2013) reforçam a ideia de que a formação de uma estrutura em árvore é essencial no sentido em que facilita a inferência probabilística, dando preferência àquelas que minimizam o tempo computacional exigido para a realização de inferência.

Antes de ser apresentado o algoritmo associado à inferência através de árvores junção, importa compreender o processo de construção de uma árvore de junção a partir de uma rede bayesiana. Neste processo de construção são usados grafos não direcionados, onde todos os vértices antecessores de cada vértice se encontram ligados por uma aresta. Estes são habitualmente designados de ***grafos morais***.

3.2.3.1 Algoritmo de criação de uma árvore de junção

Fenton e Neil (2013) resumem o algoritmo de construção de uma árvore de junção a partir de uma rede bayesiana a três etapas, nomeadamente:

1. Construção de um *grafo moral* a partir da rede bayesiana;
2. Triangulação do grafo moral (processo descrito de seguida);
3. Ligação dos *clusters*;

que passamos a apresentar.

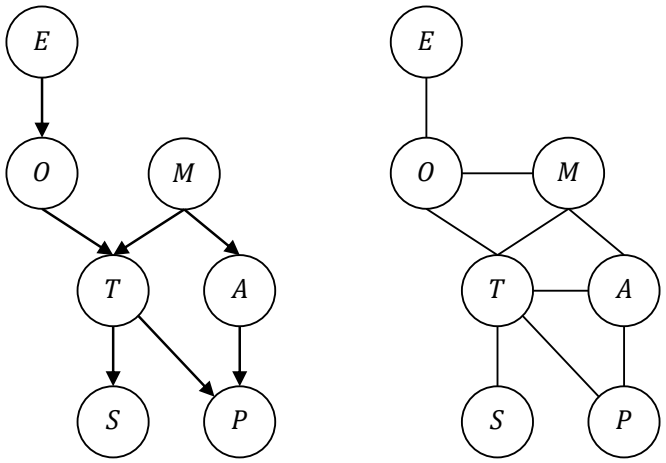
Construção de um grafo moral

A construção de um grafo moral G_M a partir do grafo G, associado a uma rede bayesiana N, passa pelos seguintes pontos:

1. Identificação em G de todos vértices antecessores de cada vértice V;

2. Adição de uma aresta não direcionada que ligue cada par de antecessores de um vértice V, caso esses não se encontrem ligados;
3. Substituição das arestas direcionadas por arestas não direcionadas.

Consideremos, a título de exemplo uma rede bayesiana representada pelo grafo (a) da Figura 3.1**Erro! A origem da referência não foi encontrada.**. Neste exemplo, existem dois pares de vértices antecessores que não se encontram ligados, que são o par O e M, antecessores do vértice T, e o par T e A, antecessores do vértice P. Se adicionarmos arestas não direcionadas entre os pares identificados, e se retiramos as direções das restantes arestas, obtemos o grafo moral (b) representado na Figura 3.1.



(a) (b)

Figura 3.1 – (a) Exemplo de rede bayesiana, com estrutura G; (b) Grafo moral G_M de G.

Triangulação de um grafo moral

Uma vez construído o grafo moral a partir de uma rede bayesiana, deve passar-se à triangulação do mesmo, etapa que tem por objetivo a identificação de subconjuntos de vértices chamados ***cluster***. Esta operação consiste nas tarefas de *identificação* e *eliminação*, sucessivas, dos vértices com o menor peso, sendo o **peso de um vértice**, recordamos, o número de arestas que é necessário adicionar entre todos os pares de vértices seus vizinhos, de modo a que o vértice e os seus vizinhos formem um **subgrafo completo**.

Exemplificando a identificação do peso de vértices a partir dos vértices o grafo (b) da Figura 3.1, tem-se que os vértices com menor peso, nomeadamente nulo, são E, S e P uma vez que, em conjunto com os seus vizinhos, estes formam já subgrafos completos, compostos respetivamente pelos conjuntos de vértices $\{E, O\}$, $\{S, T\}$ e $\{A, P, T\}$. De seguida os vértices com menor peso são os vértices M e A, ambos com peso igual a um, uma vez que é necessária

a adição de uma única aresta entre os seus vizinhos, respetivamente entre O e A, e entre M e P, de modo a formar os subgrafos completos $\{A, M, O, T\}$ e $\{A, M, P, T\}$. Por fim o vértice com maior peso é o T, com um peso igual a 7.

O processo de triangulação segue então os seguintes passos:

1. Determinação do vértice com o menor peso, identificando, se necessário, a arestas têm de ser adicionadas para garantir a formação de um subgrafo completo composto por esse vértice e os seus vizinhos;
2. Adição ao grafo das arestas identificadas no passo anterior;
3. Definição do *cluster* composto pelo vértice selecionado no passo 1 e os seus vizinhos;
4. Remoção do vértice selecionado no passo 1, e das suas arestas adjacentes;
5. Repetição dos passos 1 a 4 até se eliminarem todos os vértices.

Note-se que no passo 1 podem existir vários vértices. Nesse caso deve-se escolher arbitrariamente um desses. Note-se também que poderão existir diferentes configurações de árvores junção para uma mesma rede bayesiana.

Tomando como exemplo o grafo moral (b) da Figura 3.1, uma das concretizações possíveis para o algoritmo da triangulação é a seguinte:

1. Os vértices com o menor peso no grafo são E, S e P, tendo um peso nulo. Escolhendo E como o primeiro vértice a eliminar, dado que o seu peso é nulo, não é necessário adicionar arestas. Eliminando E, identifica-se o *cluster* EO. Desta eliminação resulta o grafo (b) da Figura ;
2. No grafo resultante, os vértices com menor peso continuam a ser S e P. Seleciona-se, por exemplo, S para eliminar, não sendo novamente necessário adicionar arestas. Ao se eliminar S identifica-se o *cluster* ST. Desta eliminação resulta o grafo (c) da Figura 3.2;
3. De seguida os vértices com menor peso, sendo este nulo, são O e P. Escolhendo eliminar O e identifica-se o *cluster* MOT. Desta eliminação resulta o grafo (d) da Figura 3.2
4. Nesta última figura os vértices M e P têm o peso mínimo, que é nulo. Eliminando M identifica-se o *cluster* AMT e resulta o grafo (e) da Figura 3.2;
5. No grafo (e) todos os vértices têm peso nulo. Optando por eliminar, por exemplo, P, identifica-se o *cluster* APT. Desta eliminação resulta o grafo (f) da Figura 3.2;

6. Novamente, neste ponto todos os vértices têm peso nulo. Eliminando, por exemplo, o vértice A, identifica-se o *cluster* AT. Desta eliminação resulta o grafo (g) da Figura 3.2;
7. Por fim, elimina-se T e identifica-se o *cluster* T.

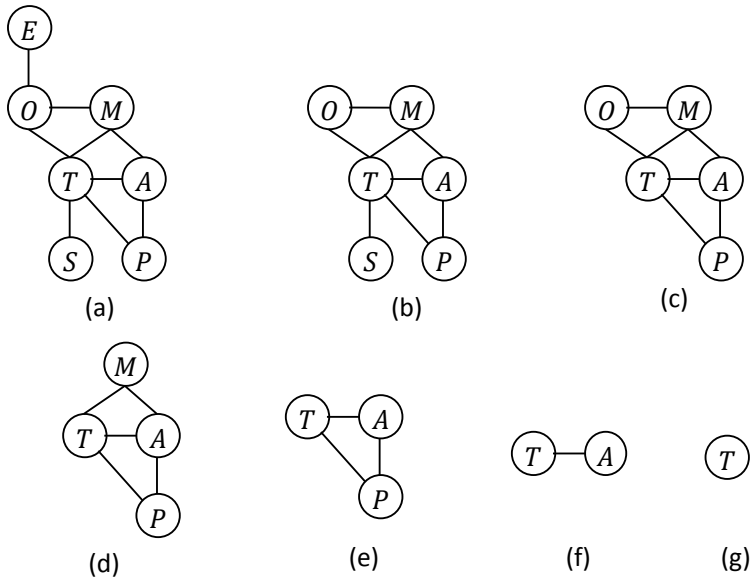


Figura 3.2 – Etapas da operação de triangulação do grafo (b) da Erro! A origem da r

Ligação dos clusters

O último passo do processo de construção de uma árvore de junção consiste na ligação dos *clusters* obtidos na etapa da triangulação seguindo os seguintes passos:

1. Partindo do conjunto de *clusters* identificados na etapa de triangulação, eliminar aqueles que não são maximais, ou seja, todos aqueles que se encontram contidos noutro *cluster* maior;
2. Para os *clusters* restantes formar uma árvore de junção, ligando pares de *clusters* com separadores que contenham as variáveis em comum. Cada árvore com *c* *clusters*, deve ter $(c-1)$ separadores.

Note-se que os vértices de uma árvore de junção são então os *clusters* identificados, sendo cada aresta associada ao separador correspondente.

Note-se ainda que a seleção dos candidatos a separadores deve passar pela identificação daqueles com o maior número de membros. No caso de empate, entre dois ou mais candidatos, deve optar-se por aquele com menor peso, ou seja, aquele em que a união dos suportes das variáveis aleatórias envolvidas tem um menor cardinal.

Na etapa de triangulação exemplificada a partir do grafo moral da Figura 3.1 foram identificados os seguintes *clusters*:

$$EO, ST, MOT, AMT, APT, AT, T. \tag{3.28}$$

Descartam-se os dois últimos *clusters*, nomeadamente AT e T, por não serem maximais, uma vez que estão contidos no *cluster* APT. Temos então que os *clusters* que vão compor a árvore de junção construída a partir do grafo da Figura 3.1 são:

$$EO, ST, MOT, AMT, APT. \tag{3.29}$$

Um possível conjunto de separadores válido para este conjunto de *clusters* é composto pelo separador O entre os *clusters* EO e MOT, o separador T entre os *clusters* ST e MOT, o separador MT entre os *clusters* MOT e AMT, o separador AT entre os *clusters* AMT e APT. Uma possível apresentação para a árvore de junção resultante encontra-se na Figura 3.3 onde se representam os separadores com retângulos e os *clusters* com elipses.

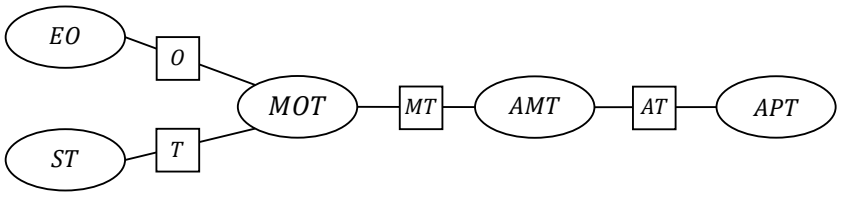


Figura 3.3 – Exemplo de árvore de junção associada à rede bayesiana da Figura 2.9.

3.2.3.2 Propagação de informação

Deste modo, uma árvore de junção T, construída a partir da rede bayesiana $N = (\boldsymbol{X}, G, \boldsymbol{P})$, consiste num par $T = (\boldsymbol{C}, \boldsymbol{S})$, composto por um conjunto $\boldsymbol{C}$, com *c clusters* C_i, e um conjunto $\boldsymbol{S}$, com $(c-1)$ separadores $S_{i,j}$, que ligam pares de *clusters* C_i e C_j, *i.e.*, $S_{i,j} = C_i \cap C_j$.

No sentido de verificação da validade destas estruturas, Koller e Friedman (2009) mostram que as árvores junção satisfazem a **propriedade da interseção corrente** (*running intersection property)*. Segundo esta propriedade, dada uma árvore de junção T, sempre que se verifica $X_i \in C_i$ e $X_i \in C_j$, então também se verifica que X_i pertence a todos os *clusters* no único caminho existente entre C_i e C_j.

O processo de inferência através de árvores junção passa pela **propagação de informação** entre *clusters* vizinhos, operações a que Kjærulff e Madsen (2008) se referem como **transmissão de mensagens** ou **absorção de informação**. Note-se que estas designações advêm do facto da transmissão de uma mensagem entre um par de *clusters* adjacentes, de C_i para C_j, pode também ser interpretada como a absorção da informação do *cluster* C_i pelo *cluster* C_j. Estas operações são feitas por etapas, de vizinho para vizinho, através a manipulação de fatores. Um algoritmo que define este processo, considerado por Kjærulff e Madsen (2008) é o algoritmo **Hugin**.

Antes iniciar o algoritmo, todas as distribuições condicionais da rede bayesiana base, ou seja, as distribuições $P\big(X_i|ant(X_i)\big)$, devem ser associadas a um, a um só, *cluster* que contenha a variável X_i e as variáveis antecessoras de X_i. Como cada distribuição é associada apenas a um *cluster*, podem existir *clusters* sem distribuições associadas. Nestes casos, deve-se atribuir a esses *clusters* vetores com 1's. Depois de associadas todas as distribuições condicionais, multiplicam-se as atribuídas ao mesmo *cluster*, formando os chamados ***potenciais iniciais*** de cada *cluster*. O fator de um *cluster* C_i será denotado por ϕ_{C_i} .

No algoritmo **Hugin**, um *cluster* é escolhido como ***cluster* raiz** e o processo de passagem de mensagens dá-se em duas fases: a **recolha de mensagens**, quando as mensagens passam sucessivamente dos *clusters* folha para o *cluster* raiz, e a **distribuição de mensagens** quando as mensagens passam sucessivamente do *cluster* raiz para os *clusters* folha. Um vértice apenas pode mandar uma mensagem para outro, quando já recebeu mensagens de todos os outros seus vizinhos, consoante o sentido de propagação de informação. Por exemplo, na árvore representada na Figura 3.3, no caso da propagação de informação estar na fase de recolha de mensagens e o *cluster* raiz ser APT, o *cluster* OMT não pode passar uma mensagem ao seu vizinho AMT sem antes receber mensagens de EO e ST.

A passagem de uma mensagem entre cada par de *clusters* adjacentes, de C_i para C_j, segue os seguintes passos, *cf.* Kjærulff e Madsen (2008):

1. Calcular o fator do separador atualizado:

$$\phi^*_{S_{i,j}} = \sum_{C_i - S_{i,j}} \phi_{C_i} \; ; \tag{3.30}$$

2. Atualizar o fator do *cluster* C_i:

$$\phi^*_{C_i} = \phi_{C_i} \frac{\phi^*_{S_{i,j}}}{\phi_{S_{i,j}}}; \tag{3.31}$$

Note-se que no passo inicial $\phi_{S_{i,j}} = 1, \forall i, j$

3. Associar o fator atualizado com o separador:

$$\phi^*_{S_{i,j}} = \phi_{S_{i,j}}. \tag{3.32}$$

Nas árvores junção a evidência não é introduzida como no algoritmo de eliminação de variáveis, ou seja, como um vetor composto por uma sequência de variáveis observadas, associadas aos respetivos estados observados $(E_1 = e_1, \dots, E_k = e_k)$. Segundo Jensen (1996), neste caso, a evidência é introduzida, como um vetor $\boldsymbol{e} = (f_1, \dots, f_m)$, onde cada f_i é um vetor designado por **vetor de descobertas** (*findings*), composto por zeros, correspondentes a estados impossíveis, e uns, correspondentes aos estados possíveis de cada variável alvo $A_i \in \{A_1, \dots, A_m\}$.

De forma resumida, o processo de propagação de evidência definido pelo algoritmo *Hugin* segue os seguintes passos:

1. Associar a cada *cluster* distribuições condicionais dos vértices da rede bayesiana que envolvem as variáveis que estão incluídas no mesmo;
2. Calcular os potenciais iniciais ϕ_{C_i} de cada *cluster* $C_i \in \boldsymbol{C}$, multiplicando as distribuições associadas;
3. Iniciar com $\phi_{S_{i,j}} = 1, \forall i, j$;
4. Incorporar a evidência na árvore de junção, multiplicando cada f_i, do vetor $\boldsymbol{e} = (f_1, \dots, f_m)$, pelo fator inicial de um *cluster*, ϕ_{C_i}, que contenha a variável A_i no seu suporte;
5. Escolher um *cluster* como ***cluster* raiz**;
6. Efetuar o processo de **recolha de informação**, passando as mensagens dos *clusters* folha até o *cluster* raiz, através dos separadores, atualizando sucessivamente os potenciais dos *clusters* $\left(\phi_{C_i} \rightarrow \phi^*_{C_i}\right)$ e dos separadores $\left(\phi_{S_{i,j}} \rightarrow \phi^*_{S_{i,j}}\right)$;
7. Efetuar o processo de **distribuição de informação**, passando as mensagens no sentido oposto, do *cluster* raiz até os *clusters* folha, atualizando novamente os potenciais dos *clusters* $\left(\phi^*_{C_i} \rightarrow \phi^{**}_{C_i}\right)$ e dos separadores $\left(\phi^*_{S_{i,j}} \rightarrow \phi^{**}_{S_{i,j}}\right)$, sucessivamente.

Após a realização de uma ronda de propagação de informação, composta pelas etapas de recolha e distribuição de informação, a árvore encontra-se em **equilíbrio**, ou, segundo

Koller e Friedman (2009), **calibrada**. Os potenciais finais associados a qualquer separador $\left(\phi_{S_{i,j}}^{**}\right)$ e *cluster* $\left(\phi_{C_i}^{**}\right)$ consistem na distribuição probabilística conjunta das variáveis contidas nos mesmos, a menos de uma constante de normalização. Assim, a estimação da distribuição de uma variável X_i pode ser feita através da seleção de um *cluster* que a contenha e da subsequente marginalização das outras variáveis.

Koller e Friedman (2009) designam os potenciais finais dos *clusters* por **crenças** (*beliefs*) dos *clusters*, e os potenciais finais dos separadores por crenças dos separadores. Deste modo, dado um par de *clusters* adjacentes C_i e C_j, este encontra-se **calibrado** se:

$$\sum_{C_i - S_{i,j}} \phi_{C_i}^{**} = \phi_{S_{i,j}}^{**} = \sum_{C_j - S_{i,j}} \phi_{C_j}^{**}. \tag{3.33}$$

Note-se que numa árvore calibrada todos os pares de *clusters* se encontram calibrados.

As árvores junção definem a distribuição conjunta das variáveis da rede bayesiana da seguinte forma*:*

$$P(\boldsymbol{X}) = \frac{\prod_i \phi_{C_i}^{**}}{\prod_{(i,j)} \phi_{S_{i,j}}^{**}}. \tag{3.34}$$

Note-se que:

$$\phi_{C_i}^{**} = \sum_{\boldsymbol{X} - C_i} P(\boldsymbol{X}). \tag{3.35}$$

A capacidade de responder a questões parciais $P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e})$ a partir de árvores junção calibradas é simples no caso das variáveis ou variáveis alvo em $\boldsymbol{A}$ estarem contidas no mesmo *cluster*. Neste caso, depois de calibrada a árvore, seleciona-se o *cluster*, ou um *cluster* (no caso de haver mais do que um), que contenha essas variáveis e normaliza-se a distribuição associada ao mesmo. No caso do *cluster* selecionado conter variáveis para além das que se pretende, essas devem ser marginalizadas da distribuição. Quando as variáveis $\boldsymbol{A}$ de $P(\boldsymbol{A}|\boldsymbol{E} = \boldsymbol{e})$ estão em *clusters* separados, Koller e Friedman (2009) referem uma abordagem *ingénua* que passa pela estruturação de uma árvore de junção em que essas variáveis fiquem no mesmo *cluster*. Contudo, esta abordagem, ao forçar a alteração da estrutura já construída, anula de certa forma a sua vantagem como ferramenta de inferência. Deste modo, os autores apresentam uma abordagem alternativa que consiste na realização de eliminação de variáveis numa árvore de junção equilibrada T.

Resumindo, este algoritmo permite o cálculo da probabilidade conjunta $P(\boldsymbol{A})$ para um subconjunto arbitrário $\boldsymbol{A}$, usando as distribuições de probabilidade de uma árvore calibrada na definição de fatores correspondentes às probabilidades condicionais em $P(\boldsymbol{X})$, e a eliminação de variáveis $\boldsymbol{O}$, em que $\boldsymbol{O} = \boldsymbol{X} - \boldsymbol{A} - \boldsymbol{E}$, no conjunto de fatores resultante. Esta abordagem permite poupança de operações relativamente à eliminação de variáveis simples pois incide apenas sobre uma porção da árvore que contém as variáveis $\boldsymbol{A}$ contidas na questão parcial.

Uma possível árvore de junção com as distribuições associadas a cada vértice da rede aos *clusters* adequados encontra-se representada na Figura 3.4.

Um exemplo de aplicação do algoritmo *Hugin* encontra-se apresentado no Anexo A, onde é introduzida a evidência $M = s, A = n$, na rede considerada na Figura 3.4.

Se a questão parcial que se deseja responder é por exemplo, $P(S, T|M = s, A = n)$, essa corresponde à crença no *cluster* C_2 resultante da propagação de informação em ambos os sentidos da árvore. Se a questão parcial tiver apenas uma variável alvo, por exemplo T, e a mesma evidência, *i.e.*, $M = s, A = n$, então para estimar $P(T|M = s, A = n)$ basta escolher um *cluster* ou separador que contenha T no seu suporte, e marginalizar as outras variáveis. Se selecionamos por exemplo C_3 ou $S_{3,4}$, para obter $P(T|M = s, A = n)$, basta realizar respetivamente $\sum_{O,M} \phi^{**}_{C_3}$ ou $\sum_{M} \phi^{**}_{S_{3,4}}$.

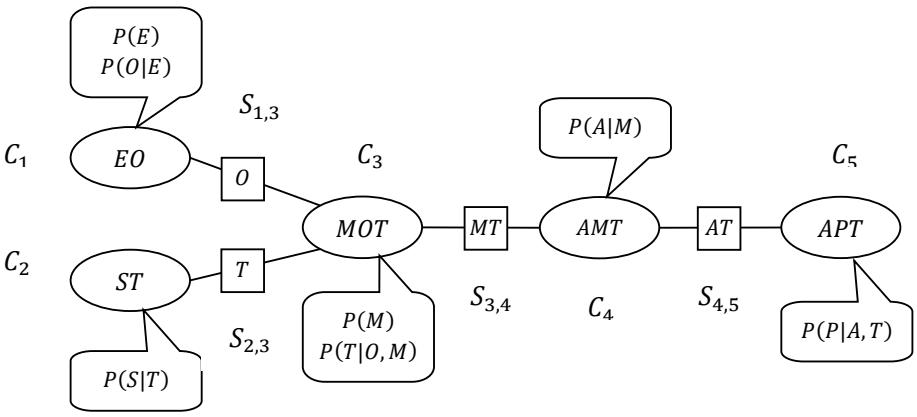


Figura 3.4 – Exemplo de árvore de junção associada à rede bayesiana da Figura 2.9, com as distribuições associadas aos *clusters*.

No caso das variáveis alvo da questão parcial estarem contidas em *clusters* separados, identifica-se, em primeiro lugar, a subárvore que as contêm. Por exemplo no caso da questão

parcial ser $P(E, P|M = s, A = n)$ a subárvore necessária para responder à mesma encontra-se identificada na Figura 3.5.

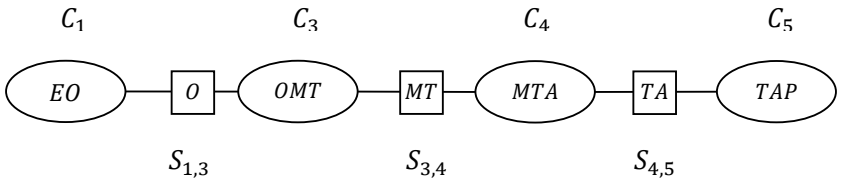


Figura 3.5 – Subárvore da árvore de junção representada na Figura 3.4.

Segundo Koller e Friedman (2009) para obter a resposta a esta questão parcial basta eliminar as variáveis não necessárias da distribuição conjunta fatorizada pelas crenças dos *clusters* e dos separadores das subárvores, *i.e.*,

$$P(E, P|M = s, A = n) = \sum_{\substack{O,M,T,A,\\ M=s,A=n}} \frac{\phi^{**}_{C_1}\phi^{**}_{C_3}\phi^{**}_{C_4}\phi^{**}_{C_5}}{\phi^{**}_{S_{1,3}}\phi^{**}_{S_{3,4}}\phi^{**}_{S_{4,5}}} = \tag{3.36}$$

$$\sum_{\substack{O,M,T,A,\\ M=s,A=n}} \frac{P(E, O|M = s, A = n)P(O, M, T|M = s, A = n)P(M, T, A|M = s, A = n)P(T, A, P|M = s, A = n)}{P(O|M = s, A = n)P(M, T|M = s, A = n)P(T, A|M = s, A = n)}.$$

3.3 Resolução de questões MAP e MPE

3.3.1 Algoritmo do produto máximo para a eliminação de variáveis

Koller e Friedman (2009) começam por apresentar a forma mais básica de resolução de questões MAP através do algoritmo de eliminação de variáveis. Este difere no ponto 5 do apresentado na página 37 aplicado às CPQ, onde se substitui a operação de multiplicação $\left(\sum_{O_i}\psi\right)$ pela de maximização $\left(\max_{O_i}\psi\right)$. Difere também pela introdução de um novo procedimento, inserido antes do passo 6 desse mesmo algoritmo, chamado de ***traceback***, da seguinte forma: dados os fatores ϕ_{X_i}onde $\{\phi_{X_i}: i = 1, \dots, k\}$:

1. Calcular a instanciação das variáveis $X_{i+1}, \dots, X_k$, $\boldsymbol{u_i} = (x^*_{i+1}, \dots, x^*_k)$, que maximiza o produto dos fatores das variáveis eliminadas antes de X_i;
2. Calcular $x^*_i = \underset{x_i}{\operatorname{argmax}}\, \phi_{X_i}(x_i, \boldsymbol{u_i})$, que corresponde à maximização da entrada correspondente no fator, relativa às anteriores escolhas $\boldsymbol{u_i}$;
3. Devolver x^*_i.

Segundo os autores, os mesmos algoritmos usados no caso das CPQ podem ser aplicados para a resolução das questões MAP. Como nas CPQ, diferentes ordens de

eliminação das variáveis podem induzir diferentes graus de complexidade de aplicação do algoritmo. O relaxamento da ordenação das variáveis a eliminar pode ser benéfico no caso das questões MAP no sentido em que, segundo os autores, mesmo o melhor resultado consistente com a ordem imposta pode ter uma complexidade maior do que uma boa ordem sem restrição, ou seja, não imposta. Estes referem ainda que mesmo para estruturas mais simples o algoritmo de eliminação de variáveis pode levar um tempo exponencial para resolver uma questão MAP. Dado que a complexidade de aplicação deste algoritmo depende da estrutura da rede, e da escolha de maximização das variáveis, a restrição da ordem de eliminação pode ser decisiva.

A título de exemplo, consideremos uma rede de pequena dimensão, nomeadamente $A \rightarrow B$, a partir da qual se deseja encontrar a instanciação que maximiza a distribuição conjunta das variáveis, ou seja, $(a^*, b^*) = \underset{A,B}{\operatorname{argmax}} P(A, B)$. Consideremos as distribuições apresentadas na Tabela 3.4 e na Tabela 3.5.

Tabela 3.4 – Distribuição $P(A)$.

A	$P(A)$
s	0,46
n	0,54

Tabela 3.5 – Distribuição $P(B|A)$.

B	A	$P(B\|A)$
s	s	0,23
n	s	0,77
s	n	0,61
n	n	0,39

Na escolha da instanciação maximizadora, independentemente da escolha do valor de A, o valor de B deve ser selecionado de acordo com a maximização de $P(B|A)$. Deste modo, observando a Tabela 3.5, temos que $\max_B P(B|A = s) = P(B = n|A = s) = 0{,}77$, e que $\max_B P(B|A = n) = P(B = s|A = n) = 0{,}61$. Assim:

$$\begin{aligned} \max_{A,B} P(A, B) &= \max_A \max_B P(A)P(B|A) \\ &= \max\{0{,}46 \times 0{,}77; 0{,}54 \times 0{,}61\} \\ &= \max\{0{,}3542; 0{,}3294\} \\ &= 0{,}3542. \end{aligned} \tag{3.37}$$

A instanciação maximizadora é então $A = s, B = n$.

Suponhamos que queremos determinar a instanciação mais provável da rede sem evidência. Os passos da eliminação de variáveis são apresentados na Tabela 3.6.

No primeiro passo $\tau_1(O) = \max_E \phi_E(E)\phi_O(O,E)$, cujo resultado equivale a $(\tau_1(O=s) = 0.98 \times 0.25 = 0.245; \tau_1(O=n) = 0.85 \times 0.75 = 0.6375)$. Continuando o processo, no último passo, o fator $\tau_7(\emptyset)$ corresponde ao valor de $\max_{E,O,M,A,T,P} P(E,O,M,A,T,P)$.

Tabela 3.6 – Passos da eliminação de variáveis para a resolução de MAP sem evidência da rede da Figura 2.9.

$Passo$	$Variável\ Eliminada$	$Fatores\ usados$	$Fator\ Intermédio$	$Fator\ Novo$
1	E	$\phi_E(E), \phi_O(O,E)$	$\psi_1(E,O)$	$\tau_1(O)$
2	O	$\tau_1(O), \phi_T(O,M,T)$	$\psi_2(O,M,T)$	$\tau_2(M,T)$
3	M	$\tau_2(M,T), \phi_M(M), \phi_A(A,M)$	$\psi_3(A,M,T)$	$\tau_3(T,A)$
4	S	$\phi_S(S,T)$	$\psi_4(S,T)$	$\tau_4(T)$
5	A	$\tau_3(T,A), \phi_P(P,A,T)$	$\psi_5(A,P,T)$	$\tau_5(T,P)$
6	T	$\tau_4(T), \tau_5(T,P)$	$\psi_6(P,T)$	$\tau_6(P)$
7	P	$\tau_6(P)$	$\psi_7(P)$	$\tau_7(\emptyset)$

Depois da eliminação segue-se a etapa de *traceback*, onde a ordem das variáveis é oposta à da eliminação. Neste exemplo, esta etapa decorre da seguinte forma:

$$\begin{aligned}
P^* &= \operatorname{argmax}_P \psi_7(P) = (P=s);\\
T^* &= \operatorname{argmax}_T \psi_6(P^*,T) = (T=s);\\
A^* &= \operatorname{argmax}_A \psi_5(A,P^*,T^*) = (A=s);\\
S^* &= \operatorname{argmax}_S \psi_4(S,T^*) = (S=s);\\
M^* &= \operatorname{argmax}_M \psi_3(A^*,M,T^*) = (M=s);\\
O^* &= \operatorname{argmax}_O \psi_2(O,M^*,T^*) = (O=n);\\
E^* &= \operatorname{argmax}_E \psi_1(E,O^*) = (E=n).
\end{aligned} \tag{3.38}$$

Temos então que:

$$\begin{aligned}
\max_{E,O,M,A,T,P} P(E,O,M,A,T,P) &= P(E=n)P(O=n|E=n)P(M=s)\\
&\quad P(T=s|M=s,O=n)P(S=s|T=s)\\
&\quad P(A=s|M=s)P(P=s|A=s,T=n)\\
&= 0{,}75 \times 0.85 \times 0.6 \times 0.6 \times 0.82 \times 0.9 \times 0.9\\
&\approx 0.152.
\end{aligned} \tag{3.39}$$

Nos passos intermédios da resolução de uma questão MPE devem ser marginalizadas as variáveis sobre as quais não se pretende inferir. Por exemplo, imaginemos que desejamos

calcular a instanciação mais provável para as variáveis T, P e A, ou seja, $\max_{A,T,P} \sum_{E,O,M,S} P(E,O,M,A,T,P)$. A eliminação e maximização das variáveis são apresentadas pelas operações:

1. Eliminar S:

$$\begin{aligned} \psi_1(S,T) &= \phi_S(S,T); \\ \tau_1(T) &= \sum_S \psi_1(S,T); \end{aligned} \tag{3.40}$$

2. Eliminar M:

$$\begin{aligned} \psi_2(A,M,O,T) &= \phi_T(M,O,T)\phi_M(M)\phi_A(A,M); \\ \tau_2(A,O,T) &= \sum_M \psi_2(A,M,O,T)\,; \end{aligned} \tag{3.41}$$

3. Eliminar O:

$$\begin{aligned} \psi_3(A,E,O,T) &= \phi_O(E,O)\tau_2(A,O,T); \\ \tau_3(A,E,T) &= \sum_O \psi_3(A,E,O,T)\,; \end{aligned} \tag{3.42}$$

4. Eliminar E:

$$\begin{aligned} \psi_4(A,E,T) &= \tau_3(A,E,T)\phi_E(E); \\ \tau_4(A,T) &= \sum_E \psi_4(A,E,T); \end{aligned} \tag{3.43}$$

5. Maximizar A:

$$\begin{aligned} \psi_5(A,P,T) &= \tau_4(A,T)\phi_P(A,P,T); \\ \tau_5(P,T) &= \max_A \psi_5(A,P,T)\,; \end{aligned} \tag{3.44}$$

6. Maximizar T:

$$\begin{aligned} \psi_6(P,T) &= \tau_5(P,T); \\ \tau_6(P) &= \max_T \psi_6(P,T)\,; \end{aligned} \tag{3.45}$$

7. Maximizar P:

$$\begin{aligned} \psi_7(P) &= \tau_6(P); \\ \tau_7(\emptyset) &= \max_P \psi_7(P). \end{aligned} \tag{3.46}$$

Depois dos passos anteriores calcula-se $P^* = \operatorname{argmax}_P \psi_7(P)$, $T^* = \operatorname{argmax}_T \psi_6(P^*,T)$ e $A^* = \operatorname{argmax}_A \psi_5(A,P^*,T^*)$.

3.3.2 Algoritmo de árvores junção

De acordo com Koller e Friedman (2009), tal como as questões parciais probabilísticas, as questões MAP podem ser resolvidas através de árvores junção com a aplicação do algoritmo *Hugin*, mas com um tipo de propagação de informação diferente da anterior, chamada **propagação-máxima** (*cf.*, *e.g.*, Jensen (1996)). A **propagação-máxima** consiste em duas etapas - a **recolha-máxima** e a **distribuição-máxima** - que se procedem de forma análoga às etapas de recolha e distribuição de informação nas árvores junção relativas CPQ. Contudo, a atualização das crenças de cada *cluster* no processo de passagem de mensagens é chamada de **calibração-máxima** e diferencia-se da anterior na operação de marginalização, que é substituída pela de maximização na atualização do fator de cada separador, da seguinte forma:

$$\phi^*_{S_{i,j}} = \max_{C_j - S_{i,j}} \phi_{C_i}. \tag{3.47}$$

A atualização dos *clusters* segue de igual modo apresentado na expressão (3.31).

Koller e Friedman (2009) referem, que tal como no anterior algoritmo Hugin, as crenças atualizadas dos *clusters* e separadores após a calibração-máxima definem a distribuição conjunta das variáveis da rede, definida pela expressão (3.34). Segundo Jensen (1996) é esta característica da calibração-máxima que permite o estabelecimento de um método para determinar as configurações de probabilidade máxima. Deste modo, após a realização da etapa calibração-máxima, se, no final, a tabela associada a cada *cluster* tiver apenas uma configuração maximal, então essa configuração faz parte da única solução da questão MAP. Contudo, se alguma tabela tiver várias configurações maximais então existem várias soluções para a questão MAP.

4 Aprendizagem de redes

Chama-se aprendizagem à tarefa de estimação e seleção de modelos. Esta tarefa pode ser feita a partir de bases de dados, sendo neste caso designada de **aprendizagem não supervisionada**, com base em testemunhos de especialistas na área do fenómeno a modelar, sendo neste caso designada de **aprendizagem supervisionada**, ou ainda uma combinação de ambas as abordagens. Kjærulff e Madsen (2008) consideram uma abordagem meramente manual, para a qual apresentam técnicas específicas, e outra que combina os dados observados com o conhecimento e experiência de peritos.

Geralmente a tarefa de indução de uma rede bayesiana a partir de uma base de dados D divide-se em dois passos, nomeadamente o da construção da estrutura G em GAD, realizado em primeiro lugar, seguido da estimação dos parâmetros Θ do conjunto de distribuições probabilísticas condicionais.

O processo de aprendizagem de redes bayesianas tem por base um conjunto de suposições que suportam o seu funcionamento e a aplicação de algoritmos. Deste modo, dada uma base de dados denotada por D, com N casos, é assumida a existência de uma distribuição associada ao processo subjacente, denotada por P_0, chamada de *distribuição probabilística subjacente do processo*. Assume-se que a base de dados pode ser representada adequadamente através de amostras de P_0, sendo o objetivo da modelação a partir de dados a estimação dessa distribuição.

Denota-se por $D = \{c^1, \dots, c^N\}$ uma base de dados D com N vetores aleatórios independentes e identicamente distribuídos, onde cada vetor $c^j = \{x_1^j, \dots, x_n^j\}, j \in \{1, \dots, N\}$ especifica uma instanciação x_i^j para cada variável $X_i \in \boldsymbol{X}$. Assume-se que cada vetor que compõe D é uma observação da distribuição P_0. O número de categorias distintas de uma variável $X \in \boldsymbol{X}$ em D será denotado por $\|X\|$.

Note-se que Scutari e Denis (2014) referem que a construção de redes exclusivamente a partir de conhecimento de especialistas codifica relações causais conhecidas e esperadas de um determinado fenómeno, resultando na construção modelos causais. Por outro lado, a aprendizagem de modelos causais a partir de dados é mais desafiante, exigindo três suposições adicionais nomeadamente: que cada variável seja condicionalmente independente das variáveis que não são seu efeito, que exista um GAD fiel à distribuição conjunta de $\boldsymbol{X}$, de

modo a que as únicas dependências na distribuição sejam as que provêm do critério de separação-*d* do GAD, e que não existam variáveis não observadas. Pearl e Verma (1991) referem que à tarefa de aprendizagem se deve aplicar a **lei de parcimónia** (ou *navalha de Occam*), que sugere que de entre um conjunto de modelos igualmente úteis, se deve preferir o mais simples. Desta condição de minimalismo resultam redes que contemplam padrões de dependência suficientes para revelar as relações causais existentes.

4.1 Aprendizagem da estrutura

Relativamente à aprendizagem da estrutura de uma rede bayesiana a partir de uma base de dados, existem três categorias principais de algoritmos, nomeadamente os algoritmos **baseados em restrições** (*constraint-based*), os **baseados numa função de pontuação** (*score-based*), e algoritmos **híbridos** que combinam processos dos anteriores.

Note-se que o número de GAD possíveis cresce muito rapidamente com o número de vértices no grafo, que pode ser calculado segundo a seguinte fórmula fornecida por Robinson (1977), onde $f(n)$ corresponde ao número de GAD possíveis com n vértices, com condição inicial $f(0) = 1$:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i). \tag{4.1}$$

Por exemplo, para $n = 1, 2, 3, 4, 5, 6$ obtemos $f(n) = 1, 3, 21, 315, 9765, 615195$.

Pearl e Verma (1991) referem que a preferência por modelos mais simples passa pelo facto de serem mais restritivos, evitando, em geral, o sobreajustamento, permitindo assim a construção de modelos mais credíveis. Nas secções 6.3 e 6.5 a 6.8 são apresentadas aplicações práticas.

4.1.1 Algoritmos baseados em restrições

Segundo Kjærulff e Madsen (2008) a abordagem dos algoritmos baseados em restrições consideram cada GAD G, de uma rede bayesiana $B = (\boldsymbol{X}, G, \boldsymbol{P})$, como um conjunto de relações de dependência e independência condicional denotadas por $\boldsymbol{M_G}$, e que podem ser lidas de acordo com o critério da separação-*d*, já apresentado na secção 2.3. Deste modo, a tarefa de aprendizagem da estrutura resume-se à identificação da estrutura que melhor codifica M_G. Esta tarefa pode ser realizada através de testes estatísticos a partir dos quais se espera conseguir identificar, na pior das hipóteses, classes de grafos equivalentes, ou seja,

conjuntos de grafos que codifiquem as mesmas relações M_G, em alternativa a uma única estrutura.

Mais formalmente, dizemos que dois modelos são estatisticamente equivalentes se, e só se, contiverem o mesmo conjunto de variáveis, e se as amostras conjuntas sobre essas variáveis não fornecerem nenhum fundamento estatístico que permita preferir um modelo em relação ao outro. Dados quaisquer dois modelos equivalentes esses têm o mesmo **grafo esqueleto**, ou seja, o mesmo grafo é obtido substituindo as arestas direcionadas por arestas não direcionadas e ambos contêm as mesmas ligações convergentes ou estruturas em V. Uma **classe de equivalência** é o conjunto maximal de grafos (do tipo GAD) com as mesmas relações de separação-*d*. Estas classes podem ser representadas por **grafos acíclicos parcialmente direcionados** ou **GAPD** que contêm simultaneamente arestas direcionadas e não direcionadas.

A introdução de conhecimento de peritos na aprendizagem da estrutura é feita mediante a especificação da presença ou ausência de arestas entre vértices, pela identificação da orientação de arestas, ou ainda uma combinação de ambos.

4.1.1.1 Testes de hipóteses estatísticas

Uma vez que os conjuntos M_G podem ser identificados através de testes de hipóteses, estes testes tomam um papel fulcral na aprendizagem baseada em restrições que se aplicam sobre cada par de variáveis, por exemplo X e Y, onde é testada em primeiro lugar a hipótese de independência marginal, com as seguintes hipóteses:

$$\begin{gathered} H_0: P(X,Y) = P(X)P(Y), i.e., X \perp Y, \\ vs. H_1: P(X,Y) \neq P(X)P(Y), \end{gathered} \tag{4.2}$$

seguida da hipótese de independência condicionada a outros subconjuntos de variáveis S_{XY} do modelo:

$$\begin{gathered} H_0: P(X,Y|\, S_{X,Y}) = P(X|\, S_{X,Y})P(Y|\, S_{X,Y}), i.e., X \perp Y|\, S_{XY}, \\ vs. H_1: P(X,Y|\, S_{X,Y}) \neq P(X|\, S_{X,Y})P(Y|\, S_{XY}). \end{gathered} \tag{4.3}$$

Note-se que de forma a evitar valores demasiados pequenos de N_{xyz}, Kjærulff e Madsen (2008) referem que se deve realizar testes de independência condicional considerando conjuntos S_{XY} com, no máximo, três variáveis.

Segundo os autores, para a realização do primeiro teste pode ser usada a estatística de teste da razão de verosimilhanças G definida por:

$$G = 2\sum_{x,y} N_{xy} \log\left(\frac{N_{xy}}{\mathbb{E}_{xy}}\right), \tag{4.4}$$

onde $\mathbb{E}_{xy} = (N_x N_y)/N$ e N_{xy} especifica o número de casos em D onde $X = x$ e $Y = y$. Para a realização do segundo teste pode ser usada a mesma estatística de teste, cuja forma de cálculo é

$$G = 2\sum_{x,y,\boldsymbol{z}} N_{xy\boldsymbol{z}} \log\left(\frac{N_{xy\boldsymbol{z}}}{\mathbb{E}_{xy\boldsymbol{z}}}\right), \tag{4.5}$$

na qual $\boldsymbol{z}$ é uma configuração de S_{XY} e $\mathbb{E}_{xy\boldsymbol{z}} = (N_{x\boldsymbol{z}} N_{y\boldsymbol{z}})/N_{\boldsymbol{z}}$.

Para além deste teste, Scutari e Denis (2014) referem ainda os testes de informação mútua (MI), que servem de medida de distância, que são equivalentes ao teste anterior, e os testes $\mathcal{X}^2$ de Pearson, cujas estatísticas de teste são, respetivamente:

$$MI = \sum_{x,y,\boldsymbol{z}} \frac{N_{xy\boldsymbol{z}}}{N} \log \frac{N_{xyz} N_{\boldsymbol{z}}}{N_{x\boldsymbol{z}} N_{y\boldsymbol{z}}}, \tag{4.6}$$

e

$$\mathcal{X}^2 = \sum_{x,y,\boldsymbol{z}} \frac{\left(n_{xy\boldsymbol{z}} - \mathbb{E}_{xy\boldsymbol{z}}\right)^2}{\mathbb{E}_{xy\boldsymbol{z}}}, \tag{4.7}$$

onde $\mathbb{E}_{xy\boldsymbol{z}}$ é o valor esperado sob a hipótese nula.

Para realização de qualquer um destes testes a hipótese nula pode ser testada com base numa distribuição assintótica $X^2_{(\|X\|-1)(\|Y\|-1)\|Z\|}$. Note-se que estes são **testes assintóticos**, na medida em que retornam valores-*p* aproximados, que no limite convergem para o verdadeiro valor-*p*.

Segundo Tsamardinos e Bourbodakis (2010), é preferível a utilização de **testes exatos**. Porém, a abordagem mais comum, chamada **abordagem condicional exata**, tem um custo computacional demasiado elevado pois exige, para a estimação do valor-*p*, o cálculo de $P\big(S_0 \geq S | (X \perp Y | S_{XY})\big)$, onde S_0 é a estatística de teste sob a hipótese nula e S é o valor observado da estatística de teste. Este cálculo implica a enumeração das entradas de todas as tabelas de contingência com as mesmas distribuições marginais. Este obstáculo pode ser contornado mediante a utilização de testes de permutação. Estes assumem que, sobre a hipótese nula de que duas variáveis X e Y são independentes, dado um subconjunto de outras variáveis, S_{XY}, qualquer permutação destas variáveis tem a mesma probabilidade de ser observada. Uma vez que o cálculo de todas as permutações possíveis equivale à enumeração de todas as tabelas possíveis com as mesmas distribuições, essa tarefa é substituída pela

amostragem aleatória do espaço das possíveis permutações para estimar $P\left(S_0 \geq S|(X \perp Y|S_{XY})\right)$. Estes testes são conhecidos como **testes de permutação de Monte Carlo** (*cf., e.g.,* Edwards (2000)). Para a realização destes testes consideremos uma base de dados $D_0 = \left\{\left(x_j, y_j, \boldsymbol{z}_j\right)\right\}_{j=1}^{N}$, D_i ($i > 0$, inteiro) a i-ésima permutação de D_0, resultante de, fixando $\boldsymbol{Z} = \boldsymbol{z}$, permutar aleatoriamente dos possíveis valores de Y. Os valores das estatísticas de teste da base de dados D_i, calculados por $\mathcal{X}^2$ ou MI, são denotados por $S(D_i)$. O ponto de partida do algoritmo de permutação é então o conjunto D_0, uma estatística de teste S e o número de permutações B, sendo este definido pelo seguinte ciclo:

1. Para cada $i = 1, \dots, B$,
   a. Permutar aleatoriamente os dados, referentes à variável Y, obtendo D_i;
   b. Calcular $S(D_i)$.
2. Obter a estimativa para o valor-*p* dada por:

$$\hat{p}_{XY|S_{XY}} = \#\{S(D_0) \leq S(D_i), i = 1, \dots, B\}/B. \tag{4.8}$$

É proposta também a abordagem *semiparamétrica* como aproximação razoável à apresentada, que é usada como forma de diminuir o número de permutações necessárias. Esta inicia com um conjunto igual ao da abordagem apresentada e é composta pelos mesmos passos, diferindo apenas na definição do valor dos graus de liberdade, denotados por gl, igual a $\overline{S(D_i)}$, a média amostral de $\left(S(D_1), \dots, S(D_B)\right)$, devolvendo $\hat{p}_{XY|S_{XY}} = 1 - F\left(S(D_0)\right)$, onde F corresponde à função cumulativa de $\mathcal{X}^2_{gl}$.

Tsamardinos e Bourbodakis (2010) referem a vantagem dos procedimentos de permutação relativamente aos assintóticos, mostrando que nos primeiros o erro de tipo I produzido corresponde ao nível de significância fixado. Mostram também que estes levam a uma aprendizagem estrutural mais robusta, ou seja, apresentando menos erros estruturais que são definidos em termos do número de arestas a mais e arestas omitidas. Estes autores referem ainda a regra de poder heurístico de alguns algoritmos, que define um limite mínimo da amostra que valide a execução de testes de independência, e a heurística do ajustamento de graus de liberdade aplicada em casos em que as amostras contêm muitas observações nulas que correspondem a *zeros estruturais*, ou seja, identificadores de uma categoria de uma variável qualitativa e não da observação de um evento.

4.1.1.2 Algoritmo PC

Scutari e Denis (2014) referem que os algoritmos baseados em restrições provêm trabalho de Pearl aplicado a modelos causais, referindo especificamente o **algoritmo IC** (do inglês, *inductive causation*) apresentado em Pearl e Verma (1991) como aquele que forneceu as bases para a aprendizagem de estruturas GAD através de testes de independência condicional. Contudo, a incapacidade de aplicação deste algoritmo na resolução de problemas práticos motivou o desenvolvimento de outros, de entre os quais se encontra o **algoritmo PC**. Segundo Le, *et al.* (2016), o nome deste algoritmo advém das iniciais dos seus autores, Peter Sprites e Clark Glymour.

O algoritmo PC é referido como a primeira aplicação prática do algoritmo IC. Spirtes *et al.* (2000) mostram que para além das suposições já referidas, se deve assumir que o número de casos da base de dados é infinitamente grande, que não existem variáveis que não tenham sido observadas, e que os testes estatísticos não têm erro. Este algoritmo de aprendizagem da estrutura encontra o GAD equivalente à estrutura GAD de P_0.

As etapas principais deste algoritmo passam ordenadamente pela:

1. Realização de testes de independência e de independência condicional entre cada par de variáveis, de modo a derivar M_G;
2. Identificação do grafo esqueleto induzido por M_G;
3. Identificação das estruturas em V;
4. Identificação das direções decorrentes.

O teste de independência condicional $X \perp Y | S_{XY}$ pode ser realizado mediante uma das hipóteses apresentadas anteriormente. Estas são testadas ordenadamente com a cardinalidade de S_{XY}, sequencialmente, igual a $0,1,2,3$. Se a hipótese nula não for rejeitada a um nível de significância $\alpha \in (0,1)$, então a pesquisa pela relação entre X e Y termina, e a relação entre as variáveis é determinada. Depois de realizados todos os testes, é obtido $M_G = M_{\perp} \cup M_{\not\perp}$.

Para uma base de dados que tivesse como resultado o grafo da Figura 2.5, teríamos, por exemplo:

$$\begin{aligned} M_{\perp} &= \{E \perp M, E \perp T | O, P \perp A | M, \dots\}, \\ M_{\not\perp} &= \{E \not\perp O, O \not\perp T, ,,\dots\}. \end{aligned} \tag{4.9}$$

O algoritmo procede com a construção do grafo esqueleto $G_E = (N_E, A_E)$, induzido pelo conjunto de relações contidas no conjunto M_G, representado na Figura Figura 4.1. Note-se que este grafo constitui uma representação mais intuitiva do que as expressões apresentadas em (4.9).

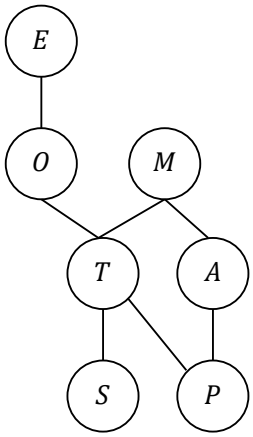


Figura 4.1– Grafo esqueleto do grafo da Figura 2.5.

Segue-se a pesquisa das estruturas em V. Esta pesquisa incide sobre todos os possíveis subconjuntos triplos de variáveis $\{X, Y, Z\}$ nos quais existem apenas dois pares de variáveis vizinhas entre si. Suponhamos, por exemplo, que são vizinhas as variáveis X e Y e as variáveis Y e Z, e que as variáveis X e Z não são vizinhas. Então sempre que $X \perp Z|S_{XZ} \in M_G$ e $Y \notin S_{XZ}$, é formada uma estrutura-v da forma: $X \longrightarrow Y \longleftarrow Z$. No fim deste passo o grafo resultante pode já ter algumas arestas direcionadas e constitui um GAPD.

No exemplo considerado identificam-se as estrutura-v: $O \longrightarrow T \longleftarrow M$ e $T \longrightarrow P \longleftarrow A$. Estas estruturas resultam de termos respetivamente $(O,T),(M,T) \in A_E$, $(O,M) \notin A_E$, e $O \perp M|S_{OM}$ quando $T \notin S_{OM}$ e $(T,P),(A,P) \in A_E$, $(A,T) \notin A_E$, e $A \perp T|S_{AT}$ quando $P \notin S_{AT}$. O grafo resultante desta etapa encontra-se representado na Figura 4.2.

Por fim, a identificação das novas direções decorrentes é feita através da aplicação de duas regras necessárias e suficientes para a obtenção da orientação maximal do GAPD. A aplicação repetida destas regras permite a obtenção das arestas orientadas em comum na classe de equivalência, garantido que não são gerados ciclos nem novas estruturas em V.

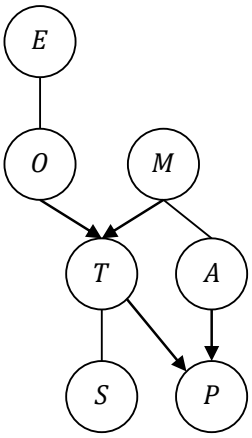


Figura 4.1 – Grafo esqueleto do grafo da Figura 2.5 com as estrutura-v identificadas.

A primeira regra define que, dada uma parcela de um grafo $X \rightarrow Y - Z$, deve-se orientar a aresta em falta da forma $X \rightarrow Y \rightarrow Z$, garantindo que não são formadas estruturas em V para além das identificadas no passo anterior. A segunda regra define que se houver no grafo um caminho direcionado de X para Y, e uma aresta entre ambas as variáveis, então deve orientar-se a mesma no sentido $X \rightarrow Y$, de modo a evitar a formação de ciclos.

No nosso exemplo, aplicando a primeira regra, direciona-se a aresta entre as variáveis T e S, resultando o grafo da Figura 4.3. Contudo, através da aplicação das regras não se consegue orientar as arestas entre E e O e entre M e A. Deste modo deve recorrer-se ao conhecimento de especialista. Por exemplo, podemos argumentar que a aresta entre E e O deve seguir de E para O, pois, geralmente, em anos de eleições políticas existem mais obras. Já a aresta entre M e A deve seguir de M para A, por ser mais provável em cenários de más condições climatéricas existirem acidentes rodoviários.

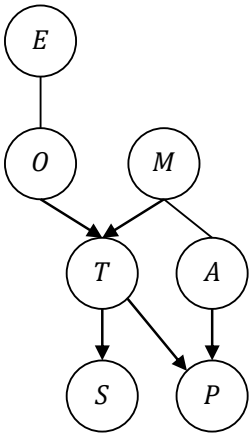


Figura 4.2 – Grafo esqueleto do grafo da Figura 2.5 com as estrutura-v identificadas e aplicação da primeira regra.

4.1.2 Algoritmos baseados em funções *score*

Os algoritmos baseados em funções *score* (pontuação) aplicam técnicas de otimização heurística a problemas de aprendizagem de estruturas de redes bayesianas, que atuam sobre um espaço de pesquisa, com o objetivo de maximizar o *score* atribuído a cada *rede candidata*. O objetivo do *score* é traduzir a adequação de cada estrutura aos dados. Estes algoritmos variam em termos do espaço de pesquisa, que podem ser estruturas, classes de estruturas equivalentes e ordens das variáveis, e na forma de pesquisa, havendo três categorias principais: a dos algoritmos **de pesquisa ávida** (*greedy search*), os **algoritmos genéticos** (*genetic*) e os de **têmpera simulada** (*simulated annealing*) (*cf., e.g.,* Scutari e Denis (2014)). Os referentes à primeira categoria exploram o espaço de estruturas de rede partindo geralmente de um cenário em que não existem arestas entre os vértices, e vão adicionando, removendo ou invertendo e arestas a cada passo, até que o *score* não aumente mais. São exemplos destes algoritmos os algoritmos de **subida de colina** (*hill climbing*) e **pesquisa tabu** (*tabu search*). Os algoritmos referentes à segunda categoria simulam a evolução natural, realizando a seleção dos modelos com o melhor *score* após a realização alternada de operações de mutação, nas quais são introduzidos alterações aleatórias, e operações que combinam as estruturas de duas redes. Por fim, os algoritmos referentes à terceira categoria efetuam uma pesquisa local estocástica, aceitando simultaneamente alterações que aumentam o *score* da rede e que diminuem com uma probabilidade inversamente proporcional à diminuição do *score*.

4.1.2.1 Algumas funções score

Quanto às medidas de *score*, Carvalho (2009) refere que estas se repartem em duas categorias: as **funções de pontuação bayesiana** (*bayesian scoring functions*) e as **baseadas na teoria da informação**.

A medida **Dirichlet bayesiana** BD (do inglês, *Bayesian Dirichlet*) proposta por Heckerman *et al.* (1995) impõe a aceitação de quatro condições. Uma delas é a condição de **amostra multinominal** que supõe que, dada uma rede B, tal como acontece nas distribuições multinominais, a observação da t-ésima instanciação dos dados é condicionalmente independente das anteriores observações. Note-se que dada uma base de dados $D = \{c^1, \dots, c^N\}$, concretização do vetor aleatório $\mathcal{D} = \{C^1, \dots, C^N\}$, uma rede B, uma variável aleatória $X_i \in \boldsymbol{X}$ e um vetor $C^t \in \mathcal{D}, t \in \{1, \dots, N\}$, esta condição pode ser expressa, através da definição da distribuição associada à rede B, $\boldsymbol{P}$, definida algebricamente da seguinte forma:

$$P\left(C_i^t = x_{ik} | C_{\text{ant}(X_i)}^t = w_{ij}, \mathcal{D}^t\right) = P\left(X_i = x_{ik} | \text{ant}(X_i) = w_{ij}\right) = \theta_{ijk}, \qquad (4.10)$$

onde $i \in \{1, \dots, n\}$, $k = 1, \dots, r_i$, onde r_i é o número de estados possíveis de X_i, e $j = 1, \dots, q_i$, sendo q_i o número de estados possíveis de $ant(X_i)$, com $\mathcal{D}^t = \{C^1, \dots, C^{t-1}\}$.

A segunda condição assume que os parâmetros $\Theta_{ij} = \{\theta_{ijk}, k = 1, \dots, r_i\}$ têm uma distribuição de Dirichlet dada uma rede G e uma base de dados D. Assim, a distribuição Θ_G associada à rede G, a sua distribuição condicionada a uma base de dados multinominal D tem uma função densidade de probabilidade dada por:

$$f(\Theta_{ij}|D, G) = c \prod_{k=1}^{r_i} \theta_{ijk}^{N'_{ijk}-1+N_{ijk}}, \tag{4.11}$$

onde $N'_{ijk} > 0$ são os parâmetros da distribuição de Dirichlet, c é uma constante de normalização e N_{ijk} é o número de observações no conjunto de dados D onde a variável X_i toma o valor x_{ik}.

A terceira hipótese refere-se à **independência dos parâmetros** que impõe que os parâmetros associados a cada variável aleatória da rede são independentes. A quarta refere-se à **modularidade dos parâmetros**, que estabelece que a probabilidade é igual para conjuntos de parâmetros associados a grafos para os quais o conjunto de vértices antecessores é comum a todas as variáveis. Sobre estas condições os autores induziram a medida BD definida por:

$$BD(B, D) = P(B) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left(\frac{\Gamma(N'_{ij})}{\Gamma(N_{ij} + N'_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ij} + N'_{ijk})}{\Gamma(N'_{ijk})} \right), \tag{4.12}$$

onde Γ é a função gama, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$ e $P(B)$ é a distribuição *a priori* da rede B. Note-se que $N'_i = \sum_{j=1}^{q_i} N'_{ij}$ são dimensões imaginárias da amostra que determinam o peso associado a cada distribuição *a priori* de X_i.

Diferentes valores de N'_{ijk} produzem diferentes medidas de score da família de BD. Por exemplo,

- Para $N'_{ij} = 1$ obtemos a medida de *score* $K2$;
- Para $N'_{ijk} = N' \times \theta_{ijk}$, sendo $N'_i = N'$ para todo o X_i, obtemos a medida **Dirichlet bayesiana de equivalência verosimilhança**, BDe (do inglês, *likelihood-equivalence Bayesian Dirichlet*).
- Para $N'_{ij} = N'/(r_i q_i)$ obtemos a medida de score **Dirichlet bayesiana equivalente uniforme** $BDeu$ (do inglês, *Bayesian Dirichlet equivalence uniform*) apresentada em Heckerman *et al.* (1995).

Relativamente às funções baseadas na teoria da informação, estas têm por base a quantidade de informação sobre D, considerando B, que é dado por:

$$\begin{aligned} L(D|B) &= -\log\big(P_B(D)\big) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\big(\theta_{ijk}\big). \end{aligned} \tag{4.13}$$

Note-se que o valor de $L(D|B)$ é mínimo quando a verosimilhança $P_B(D)$ é máxima. Isto significa que a rede que induz um código que mais comprime D é a rede que maximiza a probabilidade de observar D. Aplicando o logaritmo à verosimilhança $P_B(D)$, temos que $\log P_B(D) = -L(D|B)$, de onde resulta a **medida de log-verosimilhança** LL (do inglês, *log-likelihood*):

$$LL(D|B) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ij}}\right). \tag{4.14}$$

Segundo Carvalho (2009) esta medida facilita a ocorrência de *sobreajustamento*, favorecendo estruturas completas que não fornecem uma representação correta das suposições de independência. A adição de fatores de penalização permite a definição de melhores medidas cujas expressões seguem mais perto a abordagem da navalha de Occam. As expressões dessas medidas podem ser generalizadas por:

$$\phi(D|B) = LL(D|B) - f(N)\sum_{i=1}^{n}(r_i - 1)q_i, \tag{4.15}$$

onde $\sum_{i=1}^{n}(r_i - 1)q_i$ traduz a complexidade da rede, ou seja, o número de parâmetros necessários especificar.

Exemplos destas medidas são o **critério de informação bayesiana**, BIC (do inglês, *Bayesian information criterion*) ou MDL (do inglês, *minimum description lenght*), que consideram $f(N) = \frac{1}{2}\log(N)$. Estas são definidas por:

$$BIC(D|B) = MDL(D|B) = LL(D|B) - \frac{1}{2}\log(N)\sum_{i=1}^{n}(r_i - 1)q_i. \tag{4.16}$$

Outra medida é o **critério de informação de Akaike**, AIC (do inglês, *Akaike information criterion*), que se obtém tomando $f(N) = 1$:

$$AIC(D|B) = LL(D|B) - \sum_{i=1}^{n}(r_i - 1)q_i. \tag{4.17}$$

No capítulo 6 mostramos como podemos usar o programa *R* para aplicar as medidas de pontuação referidas.

4.1.3 Algoritmos híbridos

Estes algoritmos combinam procedimentos dos dois tipos anteriores com o objetivo de atenuar os seus aspetos menos positivos. Scutari e Denis (2014) referem que os dois algoritmos mais comuns desta categoria são o SC (do inglês, *sparse candidate*) (*cf.* Friedman *et al.* (1999)) e o $MMHC$ (do inglês, *max-min hill-climbing*) (*cf.* Tsamardinos *et al*. (2006)).

Partindo de uma estrutura candidata, estes algoritmos atuam em dois passos, nomeadamente o passo da **restrição**, no qual, para cada vértice X_i, de entre todas as restantes variáveis seleciona-se um subconjunto de variáveis C_i, candidatas a antecessoras. Este passo é seguido pelo de **maximização**, no qual se procura, entre as estruturas gráficas que traduzem as imposições definidas pelos conjuntos C_i, aquela que maximiza o *score*, dada por uma certa função. Enquanto no algoritmo SC estes passos são iterados até não haver alteração na rede ou não haver melhorias no *score* das redes testadas, no algoritmo $MMHC$ cada passo é realizado apenas uma vez.

4.2 Aprendizagem dos parâmetros

Depois de aprendida a estrutura da rede, a etapa de aprendizagem seguinte passa pela estimação dos parâmetros da mesma. Scutari e Denis (2014) referem que a tarefa de estimar e atualizar os parâmetros da distribuição global $P(\boldsymbol{X})$, onde $\boldsymbol{X}$ corresponde ao conjunto de variáveis que compõe a rede, pode ser bastante simplificada se for dividida nas distribuições locais $P\big(X_i|\mathrm{ant}(X_i)\big)$.

Existem várias hipóteses de algoritmos para estimação de parâmetros. Segundo Scutari e Denis (2014) os mais usuais dividem-se em duas categorias, sendo essas a estimação de máxima verosimilhança e a estimação bayesiana. Estes autores salientam o facto de o algoritmo utilizado para a aprendizagem da estrutura não restringir a escolha do algoritmo utilizado na aprendizagem dos parâmetros. Note-se que apesar da utilização de distribuições locais poder auxiliar na simplificação da estimação de parâmetros, esta tarefa pode ainda ser de difícil execução.

A estimação bayesiana dos parâmetros considera uma distribuição uniforme *a priori* para cada distribuição a ser estimada. Nesta estimação consideraremos também um argumento opcional, chamado *dimensão da amostra imaginária* ou iss (do inglês, *imaginary*

sample size), que permite atribuir pesos à distribuição *a priori*, e à distribuição *a posteriori*, estimada a partir dos dados disponíveis. Este peso que suporta a distribuição *a priori*, consiste na dimensão de uma amostra imaginária sendo dividida pelo número de entradas na tabela de probabilidade condicional e usado para calcular a estimativa posterior como uma média ponderada com as frequências empíricas.

Tomemos por exemplo as variáveis E e O da rede da Figura 2.9, com os estados respetivos *sim* (s) e *não* (n), e suponhamos que temos uma amostra de dimensão N. As estimativas das probabilidades, dadas pelas frequências empíricas, são:

$$\hat{p}(O = n, E = s) = \frac{nº\ de\ observações\ com\ O = n\ e\ E = s}{N}, \tag{4.18}$$

e

$$\hat{p}(E = s) = \frac{nº\ de\ observações\ com\ \ E = s}{N}. \tag{4.19}$$

As respetivas probabilidades *a priori* são dadas por:

$$\pi(O = n, E = s) = \frac{1}{nº\ de\ possíveis\ estados\ de\ O\ \times nº\ de\ possíveis\ estados\ de\ E} = 1/4\ , \tag{4.20}$$

e

$$\pi(E = s) = \frac{1}{nº\ de\ possíveis\ estados\ de\ E} = \frac{1}{2}. \tag{4.21}$$

Temos então que as probabilidades estimadas, ajustadas à dimensão da amostra imaginária são dadas por:

$$\hat{P}(O = n, E = s) = \frac{iss}{N + iss}\pi(O = n, E = s) + \frac{N}{N + iss}\hat{p}(O = n, E = s), \tag{4.22}$$

e

$$\hat{P}(E = s) = \frac{iss}{N + iss}\pi(E = s) + \frac{N}{N + iss}\hat{p}(E = s). \tag{4.23}$$

A estimativa da probabilidade posterior $\hat{P}(O = n|E = s)$, é dada por:

$$\hat{P}(O = n|E = s) = \frac{\hat{P}(O = n, E = s)}{\hat{P}(E = s)}. \tag{4.24}$$

Segundo Scutari e Denis (2014) o valor de iss deve ser relativamente pequeno, variando entre 1 e 15, de modo a que a distribuição *a priori* seja dominada pelos dados. Note-se que valores maiores de iss resultam em distribuições mais próximas da distribuição uniforme.

Os autores referem que as probabilidades estimadas por este método estão mais afastadas dos valores extremos 0 e 1 do que as estimadas pelo método da máxima verosimilhança, e que esta particularidade é vantajosa em vários aspetos. Um desses aspetos é o facto de assegurar as condições para a aplicabilidade dos métodos de estimação de modelos e de inferência, nomeadamente por não gerar tabelas com muitos valores nulos. Há ainda a vantagem de as estimativas *a posteriori* serem mais robustas, resultando em redes com maior poder preditivo.

Nas secções 6.3 e 6.4 são mostradas formas práticas de estimação de parâmetros para uma rede bayesiana no programa *R*.

5 Análise de conflitos

As redes bayesianas são meras aproximações dos fenómenos reais. Contudo, os resultados da sua aplicação prática devem ser consistentes com um conjunto de suposições acerca do funcionamento da realidade modelada. Uma análise de conflitos tem por objetivo a avaliação do funcionamento dos modelos construídos, verificando a sua consistência e garantindo a obtenção de resultados fiáveis. A evidência introduzida para inferência pode estar em conflito com o modelo. Esta análise pode ser feita com base na evidência ou em hipóteses.

5.1 Análise de conflitos motivada pela evidência

A análise com base na evidência é aplicada para identificar possíveis conflitos na evidência ou entre a evidência e o modelo, cenário no qual se pode concluir que o modelo pode ser fraco relativamente à evidência.

Considerando um vetor de evidência $\boldsymbol{E} = \boldsymbol{e}$, a medida de conflito considerada baseia-se na veracidade da seguinte igualdade:

$$P\big(E_i = e_i, E_j = e_j\big) = P\big(E_i = e_i | E_j = e_j\big)P\big(E_j = e_j\big), \forall i, j, \tag{5.1}$$

e na aceitação da suposição de que, no caso da evidência ter um comportamento "normal", se verifica, para quaisquer peças de evidência $E_i = e_i$ e $E_j = e_j$, que:

$$\begin{aligned} &P\big(E_i = e_i | E_j = e_j\big) > P(E_i = e_i) \\ \Leftrightarrow\ &P\big(E_i = e_i, E_j = e_j\big) > P(E_i = e_i)P\big(E_j = e_j\big). \end{aligned} \tag{5.2}$$

A medida de conflito indica a existência de um possível conflito quando a distribuição conjunta da evidência é menor do que o produto das probabilidades individuais de cada peça, *i.e.*, neste sentido, existe conflito entre as peças de evidência $E_i = e_i$ e $E_j = e_j$ quando estas estão negativamente correlacionadas, no sentido em que

$$\frac{P(E_i = e_i)P\big(E_j = e_j\big)}{P\big(E_i = e_i, E_j = e_j\big)} > 1 \Leftrightarrow \log\frac{P(E_i = e_i)P\big(E_j = e_j\big)}{P\big(E_i = e_i, E_j = e_j\big)} > 0. \tag{5.3}$$

A medida de conflito entre as duas peças de evidência, $E_i = e_i$ e $E_j = e_j$, é definida por:

$$\mathrm{conf}(E_i, E_j) = \log \frac{P(E_i = e_i)P(E_j = e_j)}{P(E_i = e_i, E_j = e_j)}. \quad (5.4)$$

Podemos generalizar a expressão da medida de conflito parcial acima apresentada para uma medida de conflito global, incluindo todas as k peças de evidência de $\boldsymbol{E} = \boldsymbol{e}$:

$$\mathrm{conf}(\boldsymbol{E} = \boldsymbol{e}) = \mathrm{conf}(\{E_1 = e_1, \dots, E_k = e_k\}) = \log \frac{\prod_{i=1}^{n} P(E_i = e_i)}{P(\boldsymbol{E} = \boldsymbol{e})}. \quad (5.5)$$

Um valor positivo de $conf(\boldsymbol{E} = \boldsymbol{e})$ é indicativo da existência de conflito.

Por exemplo, assumindo a rede bayesiana da Figura 2.9, e o conjunto de evidência $\{P = s, S = n, M = n\}$, arredondando os resultados a três casas decimais, temos que[4]:

$$\begin{aligned} \mathrm{conf}\{P = s, S = n, M = n\} &= \log \frac{P(\mathrm{P} = s)P(\mathrm{S} = n)P(M = n)}{P(P = s, S = n, M = n)} \\ &\approx \log \frac{0.769 \times 0.367 \times 0.400}{0.101} \approx 0.114 > 0. \end{aligned} \quad (5.6)$$

Logo a medida calculada é indicadora de conflito na evidência em causa. Note-se que todos cálculos apresentados nesta secção foram realizados com recurso ao programa R, sendo o código apresentado no Anexo B.

5.1.1 Localização de conflitos

No caso de ser detetada a existência de conflito pela medida de **conflito global**, *i.e.*, $\mathrm{conf}(\boldsymbol{E} = \boldsymbol{e})$, como no exemplo, pode ser determinada a origem do conflito. Esta determinação é facilitada pela propriedade de partição do conflito global. Sendo $\{\boldsymbol{E} = \boldsymbol{e}\} = \{\boldsymbol{E_i} = \boldsymbol{e_i}, \boldsymbol{E_j} = \boldsymbol{e_j}\}$ (note-se que $\boldsymbol{E_i} = \boldsymbol{e_i}$ e $\boldsymbol{E_j} = \boldsymbol{e_j}$ representam, cada um deles conjuntos de evidência), esta propriedade define o conflito global através das soma dos **conflitos locais** *i.e.*, $conf\{\boldsymbol{E_i} = \boldsymbol{e_i}, \boldsymbol{E_j} = \boldsymbol{e_j}\}$, com os **parciais**, $\mathrm{conf}(\boldsymbol{E_i} = \boldsymbol{e_i})$ e $\mathrm{conf}(\boldsymbol{E_j} = \boldsymbol{e_j})$. Temos então:

$$\mathrm{conf}(\boldsymbol{E} = \boldsymbol{e}) = \mathrm{conf}\{\boldsymbol{E_i} = \boldsymbol{e_i}, \boldsymbol{E_j} = \boldsymbol{e_j}\} + \mathrm{conf}(\boldsymbol{E_i} = \boldsymbol{e_i}) + \mathrm{conf}(\boldsymbol{E_j} = \boldsymbol{e_j}), \quad (5.7)$$

[4] Os cálculos deste capítulo foram efetuados com recurso ao programa R, sendo os respetivos comandos apresentados no Anexo B.

onde, se convenciona que os conflitos locais valem zero, caso a dimensão dos vetores envolvidos seja um. Por exemplo, sendo $\boldsymbol{E} = \boldsymbol{e} = (\boldsymbol{E_1}, \boldsymbol{E_2})$ e $\boldsymbol{E_1} = (E_{11}, E_{12})$ e $\boldsymbol{E_2} = E_2$ temos:

$$\begin{aligned} \text{conf}(\boldsymbol{E} = \boldsymbol{e}) &= conf\{E_{11} = e_{11}, E_{12} = e_{12}, E_2 = e_2\} \\ &= \log \frac{P(\,E_{11} = e_{11})P(\,E_{12} = e_{12})P(\,E_2 = e_2)}{P(E_{11} = e_{11}, E_{12} = e_{12}, E_2 = e_2)} \\ &= \log \frac{P(\,E_{11} = e_{11}, E_{12} = e_{12})P(\,E_2 = e_2)}{P(E_{11} = e_{11}, E_{12} = e_{12}, E_2 = e_2)} \frac{P(\,E_{11} = e_{11})P(\,E_{12} = e_{12})}{P(\,E_{11} = e_{11}, E_{12} = e_{12})} \\ &= \text{conf}(\boldsymbol{E_1}, E_2) + \text{conf}(\boldsymbol{E_1}). \end{aligned} \tag{5.8}$$

Deste modo, a identificação da origem do conflito passa pela deteção de conflitos parciais, feita através da aplicação da fórmula da medida de conflito a todos os subconjuntos evidência possíveis gerados a partir do conjunto original. Esta análise consiste em calcular $\text{conf}(\boldsymbol{E}'), \forall \boldsymbol{E}' \subseteq \boldsymbol{E}$. Todas as peças de evidência contidas nos subconjuntos $\boldsymbol{E}'$ cujos valores $\text{conf}(\boldsymbol{E}' = \boldsymbol{e}')$ são positivos correspondem a fontes de conflito.

No exemplo em questão temos os seguintes conflitos parciais:

$$\text{conf}\{P = s, S = n\} = \log \frac{P(P = s)P(S = n)}{P(P = s, S = n)} \approx \log \frac{0.769 \times 0.367}{0.232} \approx 0.196, \tag{5.9}$$

$$\text{conf}\{P = s, M = n\} = \log \frac{P(P = s)P(M = n)}{P(P = s, M = n)} \approx \log \frac{0.769 * 0.400}{0.293} \approx 0.048, \tag{5.10}$$

$$\text{conf}\{S = n, M = n\} = \log \frac{P(S = n)P(M = n)}{P(S = n, M = n)} \approx \log \frac{0.367 \times 0.4}{0.167} \approx -0.128. \tag{5.11}$$

Conclui-se dos resultados anteriores que apenas o subconjunto de evidência $\{S = n, M = n\}$ não constitui um conflito local, dado o seu resultado negativo, ao contrário dos outros dois, *i.e.*, dos subconjuntos de evidência $\{P = s, S = n\}$ e $\{P = s, M = n\}$.

Calculando os conflitos locais temos:

$$\begin{aligned} \text{conf}(\{P = s, S = n\}, M = n) &= \log \frac{P(P = s, S = n)P(M = n)}{P(P = s, S = n, M = n)} \\ &\approx \log \frac{0.232 \times 0.400}{0.101} \approx -0.081, \end{aligned} \tag{5.12}$$

$$\begin{aligned} \text{conf}(\{P = s, M = n\}, S = n) &= \log \frac{P(P = s, M = n)P(S = n)}{P(P = s, S = n, M = n)} \\ &\approx \log \frac{0.293 \times 0.367}{0.101} \approx 0.066, \end{aligned} \tag{5.13}$$

$$\begin{aligned}\text{conf}(\{S=n, M=n\}, P=s) &= log\frac{P(S=n, M=n)P(P=s)}{P(P=s, S=n, M=n)} \\ &\approx \log\frac{0.167\times 0.769}{0.101} \approx 0.242.\end{aligned} \tag{5.14}$$

Somando os valores dos conflitos locais aos parciais, como já esperado, atendendo à fórmula (5.7) temos:

$$\begin{aligned}\text{conf}(\boldsymbol{E}=\boldsymbol{e}) &= conf\{P=s, S=n\} + conf(\{P=s, S=n\}, M=n) \\ &\approx 0.196 - 0.081 = 0.114,\end{aligned} \tag{5.15}$$

$$\begin{aligned}\text{conf}(\boldsymbol{E}=\boldsymbol{e}) &= conf\{P=s, M=n\} + conf(\{P=s, M=n\}, S=n) \\ &\approx 0.048 + 0.066 = 0.114,\end{aligned} \tag{5.16}$$

$$\begin{aligned}\text{conf}(\boldsymbol{E}=\boldsymbol{e}) &= conf\{M=n, S=n\} + conf(\{M=n, S=n\}, P=s) \\ &\approx -0.128 + 0.242 = 0.114.\end{aligned} \tag{5.17}$$

5.1.2 Resolução de conflitos

Um possível indicador da existência de conflitos é a verificação de evidência em conjuntos descoberta ou de evidência com probabilidade baixa, conhecidos por **casos raros.** Sendo $\boldsymbol{E}=\boldsymbol{e}$ um conjunto indicador de conflito, ou seja, sendo $\boldsymbol{E}=\boldsymbol{e}$ tal que $\text{conf}(\boldsymbol{E}=\boldsymbol{e}) > 0$, pode haver uma hipótese $H=h$ que pode explicar as descobertas, no sentido em que $conf(\boldsymbol{E}=\boldsymbol{e} \cup \{H=h\}) \leq 0$. Nesse caso é provável que $\boldsymbol{E}=\boldsymbol{e}$ seja um caso raro e que $H=h$ possa explicar o conflito. A medida de conflito conjunta da evidência $\boldsymbol{E}=\boldsymbol{e}$ com $H=h$ pode ser calculada segundo a seguinte expressão:

$$\text{conf}(\boldsymbol{E}=\boldsymbol{e} \cup \{H=h\}) = \text{conf}(\boldsymbol{E}=\boldsymbol{e}) + \log\frac{P(H=h)}{P(H=h|\boldsymbol{E}=\boldsymbol{e})}. \tag{5.18}$$

Pois,

$$\begin{aligned}\text{conf}(\boldsymbol{E}=\boldsymbol{e} \cup \{H=h\}) &= \log\frac{P(H=h)\prod_{i=1}^{n}P(E_i=e_i)}{P(\boldsymbol{E}=\boldsymbol{e}, H=h)} \\ &= \log\frac{\prod_{i=1}^{n}P(E_i=e_i)}{P(\boldsymbol{E}=\boldsymbol{e})}\frac{P(H=h)P(\boldsymbol{E}=\boldsymbol{e})}{P(\boldsymbol{E}=\boldsymbol{e}, H=h)} \\ &= conf(\boldsymbol{E}=\boldsymbol{e}) + \log\frac{P(H=h)}{P(H=h|\boldsymbol{E}=\boldsymbol{e})}.\end{aligned} \tag{5.19}$$

Tomando os estados das variáveis A e T como hipóteses, temos:

$$\begin{aligned}\text{conf}(\boldsymbol{E}=\boldsymbol{e} \cup \{A=s\}) &= \text{conf}(\boldsymbol{E}=\boldsymbol{e}) + \log\frac{P(A=s)}{P(A=s|\boldsymbol{E}=\boldsymbol{e})} \\ &\approx 0.114 + \log\frac{0.744}{0.699} \approx 0.114 + 0.062 = 0.176 > 0,\end{aligned} \tag{5.20}$$

$$\text{conf}(\boldsymbol{E}=\boldsymbol{e} \cup \{A=n\}) = \text{conf}(\boldsymbol{E}=\boldsymbol{e}) + \log\frac{P(A=n)}{P(A=n|\boldsymbol{E}=\boldsymbol{e})} \tag{5.21}$$

$$\approx 0.114 + \log\frac{0.256}{0.301} \approx 0.114 - 0.160 = -0.046 < 0,$$

$$\begin{aligned} \text{conf}(\boldsymbol{E} = \boldsymbol{e} \cup \{T = s\}) &= \text{conf}(\boldsymbol{E} = \boldsymbol{e}) + \log\frac{P(T = s)}{P(T = s|\boldsymbol{E} = \boldsymbol{e})} \\ &\approx 0.114 + \log\frac{0.689}{0.208} \approx 0.114 + 1.195 = 1.309 > 0, \end{aligned} \tag{5.22}$$

$$\begin{aligned} \text{conf}(\boldsymbol{E} = \boldsymbol{e} \cup \{T = n\}) &= \text{conf}(\boldsymbol{E} = \boldsymbol{e}) + \log\frac{P(T = n)}{P(T = n|\boldsymbol{E} = \boldsymbol{e})} \\ &\approx 0.114 + \log\frac{0.314}{0.792} \approx 0.114 - 0.926 = -0.811 < 0. \end{aligned} \tag{5.23}$$

Conclui-se então que o conflito pode ser explicado por dois casos raros, um no qual não há nenhum acidente rodoviário, *i.e.*, $conf(\boldsymbol{E} = \boldsymbol{e} \cup \{A = n\})$, e outro no qual não há trânsito congestionado, *i.e.*, $conf(\boldsymbol{E} = \boldsymbol{e} \cup \{T = n\})$, e mesmo assim há polícia na estrada e não há sinalização de trânsito congestionado nem más condições climatéricas.

5.2 Análise de conflitos motivada pelas hipóteses

Neste tipo de análise pretende-se investigar o impacto de um vetor de evidência na probabilidade de uma hipótese a estudar, ou seja, pretende-se identificar as peças de evidência que entram em conflito com o conjunto de evidência total. A medida utilizada nesta análise fornece um valor que especifica o custo de omitir uma única peça de evidência. O **custo de omissão** de $E_i = e_i$ é definido pela seguinte expressão:

$$\begin{aligned} &c\big(P(\boldsymbol{X}|\boldsymbol{E} = \boldsymbol{e}), P(\boldsymbol{X}|\boldsymbol{E} = \boldsymbol{e} - \{E_i = e_i\})\big) \\ &= \sum_{x \in sup(\boldsymbol{X})} P(\boldsymbol{X} = x|\boldsymbol{E} = \boldsymbol{e}) \log\frac{P(\boldsymbol{X} = x|\boldsymbol{E} = \boldsymbol{e})}{P(\boldsymbol{X} = x|\boldsymbol{E} = \boldsymbol{e} \backslash \{E_i = e_i\})}, \end{aligned} \tag{5.24}$$

(*cf.* Kjærulff e Madsen (2008)).

Sendo H uma variável hipótese com o suporte $\sup(H) = \{h_1, \dots, h_m\}$, o impacto de uma descoberta $\{E_i = e_i\} \in \boldsymbol{E} = \boldsymbol{e}$ numa hipótese $h_i \in \sup(H)$ é determinado pela comparação de $P(H = h_i)$ com $P(H = h_i|\boldsymbol{E} = \boldsymbol{e})$ e $P(H = h_i|\boldsymbol{E} = \boldsymbol{e} - \{E_i = e_i\})$. Deste modo, conseguimos identificar as descobertas que geram conflitos na probabilidade da hipótese comparado com o impacto no conjunto de evidência completo.

Tabela 5.1 – Tabela que permite analisar o impacto de cada peça de evidência nas hipóteses consideradas.

Hipótese	$P(H = h)$	$P(H = h\|\boldsymbol{E} = \boldsymbol{e})$	$P(H = h\|\boldsymbol{E} = \boldsymbol{e} - \{P = s\})$	$P(H = h\|\boldsymbol{E} = \boldsymbol{e} - \{S = n\})$	$P(H = h\|\boldsymbol{E} = \boldsymbol{e} - \{M = n\})$
$A = s$	0.744000	0.6994635	0.630000	0.6872604	0.7906106
$A = n$	0.256000	0.3005365	0.370000	0.3127396	0.2093894
$T = s$	0.685975	0.2076201	0.1504783	0.7134013	0.2527331
$T = n$	0.314025	0.7923799	0.8495217	0.2865987	0.7472669

Analisando os valores da tabela anterior, e comparando os valores das probabilidades da variável A condicionada à evidência completa com aquelas em que são omitidas descobertas, concluímos que as descobertas $P = s$ e $S = n$ agem a favor a hipótese $A = s$, uma vez que os valores das probabilidades condicionadas são superiores quando estas descobertas são incluídas na evidência. Por outro lado, a descoberta $M = n$ age contra desta hipótese uma vez que a probabilidade condicionada é superior quando esta é excluída da evidência. Uma vez que a variável A é binária, as mesmas descobertas têm o efeito oposto sobre a hipótese $A = n$.

Analogamente, quanto à variável T, constata-se que a descoberta $P = s$ age a favor da hipótese $T = s$ e as descobertas $S = n$ e $M = n$ agem contra esta hipótese. Igualmente por ser binária, os efeitos das descobertas são opostos relativamente à hipótese $T = n$.

Note-se que o cálculo do custo de omissão é mais útil quando estamos perante variáveis com suportes maiores. Aplicando a fórmula ao exemplo temos:

$$\begin{aligned} & c\big(P(A|\boldsymbol{E} = \boldsymbol{e}), P(A|\boldsymbol{E} = \boldsymbol{e}\backslash\{P = s\})\big) \\ & \approx 0.699 \times \log\frac{0.699}{0.63} + 0.301 \times \log\frac{0.301}{0.37} \approx 0.0107, \end{aligned} \tag{5.25}$$

e

$$c\big(P(A|\boldsymbol{E} = \boldsymbol{e}), P(A|\boldsymbol{E} = \boldsymbol{e}\backslash\{S = n\})\big) \approx 0.0003, \tag{5.26}$$

$$c\big(P(A|\boldsymbol{E} = \boldsymbol{e}), P(A|\boldsymbol{E} = \boldsymbol{e}\backslash\{M = n\})\big) \approx 0.0230, \tag{5.27}$$

$$c\big(P(T|\boldsymbol{E} = \boldsymbol{e}), P(T|\boldsymbol{E} = \boldsymbol{e}\backslash\{P = s\})\big) \approx 0.0120, \tag{5.28}$$

$$c\big(P(T|\boldsymbol{E} = \boldsymbol{e}), P(T|\boldsymbol{E} = \boldsymbol{e}\backslash\{S = n\})\big) \approx 0.5495, \tag{5.29}$$

$$c\big(P(T|\boldsymbol{E} = \boldsymbol{e}), P(T|\boldsymbol{E} = \boldsymbol{e}\backslash\{M = n\})\big) \approx 0.0056. \tag{5.30}$$

Qualquer um destes valores é positivo, indicando custos de omissão de evidência.

6 Aplicações práticas no programa R

Existem várias ferramentas informáticas que permitem trabalhar com redes bayesianas. São exemplos destes os programas *Hugin*, *BayesianLab*, *GeNIe* e *R*. Ao contrário dos primeiros programas referidos, as funcionalidades do programa R não se resumem a estas redes. Este programa trabalha com pacotes (do inglês, *packages*) que abrangem uma variedade de funcionalidades. Esta dissertação dedica este último capítulo à apresentação sucinta de algumas funcionalidades deste programa no que toca às redes bayesianas.

Scutari e Denis (2014) referem que o crescente uso da linguagem R no contexto bayesiano advém da também crescente variedade de áreas que utilizam estas redes. Apesar da desvantagem de esta funcionar através de comandos, com os quais utilizadores sem conhecimentos de programação informática podem não estar habituados, este programa é bastante versátil em termos de aplicações no contexto das redes bayesianas, tendo uma elevada propensão à inclusão de novos métodos estatísticos. Existem vários *packages* que permitem o manuseamento de redes bayesianas entre os quais se encontram os seguintes: `bnlearn`, `deal`, `catnet`, `pcalg`, `gRain`, `gRbase` e `rbmn`. Estes possuem diferentes funcionalidades como a construção e a manipulação de estruturas de redes bayesianas discretas, contínuas e mistas, a inserção dos parâmetros, a realização de testes de independência condicional, a aplicação de algoritmos de aprendizagem baseados em restrições, em funções *score* e híbridos, a estimação de parâmetros, a previsão e a realização de inferência exata e aproximada. De seguida são apresentados alguns dos comandos básicos mais relevantes do *package* `bnlearn` que permitem aplicar os conceitos apresentados das secções anteriores. As versões do programa R e do *package* `bnlearn` usadas no presente capítulo são a 3.6.1 e a 4.4.1, respetivamente.

6.1 Construção de uma rede bayesiana conhecida

Depois de aberto o programa e instalado o *package*, este deve chamado:

```
> library(bnlearn)
```

Para, por exemplo, criar o GAD da Figura 2.9, definem-se primeiro os vértices correspondentes às variáveis, criando um GAD "vazio", uma vez que não contém de arestas, com o seguinte comando:

```
> dag<-empty.graph(nodes = c("E","O","M","T","S","A","P"))
```

Depois definem-se as arestas:

```
> dag<-set.arc(dag,from="E",to="O")

[…]

> dag<-set.arc(dag,from="T",to="P")
```

O comando `modelstring` apresenta as distribuições condicionais que devem ser definidas ou estimadas para as variáveis da rede:

```
> modelstring(dag)
[1] "[E][M][O|E][A|M][T|O:M][S|T][P|T:A]"
```

Antes da associação das distribuições à rede definem-se os suportes das variáveis, que nestes casos se resumem aos estados *sim* e *não*:

```
> E.lv<-c("sim","não")

[…]

> P.lv<-c("sim","não")
```

De seguida associam-se as distribuições às variáveis, segundo as apresentadas na Figura 2.9:

```
> E.prob<-array(c(0.25,0.75),dim=2,dimnames=list(E=E.lv))

> O.prob<-
array(c(0.98,0.02,0.15,0.85),dim=c(2,2),dimnames=list(O=O.lv,E=E.lv))

[…]

>P.prob<-
array(c(0.9,0.1,0.63,0.37,0.72,0.28,0.45,0.55),dim=c(2,2,2),dimnames=l
ist(P=P.lv,T=T.lv,A=A.lv))
```

Note-se que para a visualização da distribuição basta escrever o nome, por exemplo:

```
> O.prob
     E
O      sim  não
  sim 0.98 0.15
  não 0.02 0.85
```

Para facilitar a definição da rede bayesiana agrupam-se as distribuições num objeto, chamado, neste caso, de `cpt`:

```
>cpt<-
list(E=E.prob,O=O.prob,M=M.prob,T=T.prob,S=S.prob,A=A.prob,P=P.prob)
```

Por fim, as distribuições guardadas no objeto `cpt` juntam-se à estrutura GAD, guardada no objeto chamado `dag`, de modo a definir a rede bayesiana num objeto da classe `bn.fit`, chamado `bn`:

```
bn<-custom.fit(dag,cpt)
```

Para saber o número de parâmetros da rede pode usar-se o comando `nparams`:

```
> nparams(bn)
[1] 16
```

Para visualizar a distribuição de uma variável da rede, por exemplo a variável M, podemos correr o seguinte código:

```
> bn$M
  Parameters of node M (multinomial distribution)

Conditional probability table:
 M
sim não
0.6 0.4
```

6.2 Geração de imagens

O programa permite a geração de imagens, permitindo, por exemplo, a visualização das distribuições em tabelas:

```
> bn.fit.dotplot(bn$O, xlab = "Estados de E",
+                ylab = "Estados de O", main = "Probabilidades
condicionais P(O|E)")
```

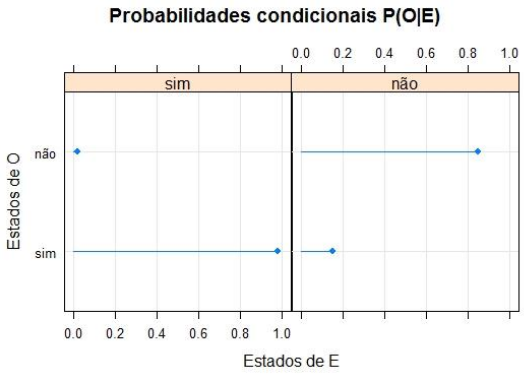


Figura 6.1 – Tabela da distribuição condicional associada à variável O gerada com a função `bn.fit.dotplot` do *package* `bnlearn`.

Para desenhar o GAD da Figura 2.9 pode executar-se a seguinte linha de código:

```
> graphviz.plot(bn, layout = "dot")
```

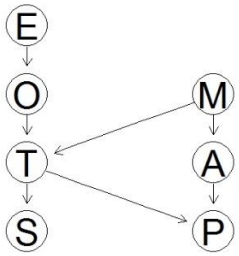


Figura 6.2 – Grafo gerado com a função `graphviz.plot` do *package* `bnlearn`.

Já o desenho da rede com as distribuições de cada variável pode ser obtido do seguinte modo:

```
> graphviz.chart(bn, layout = "dot",type="barprob",draw.levels =
TRUE,grid=TRUE)
```

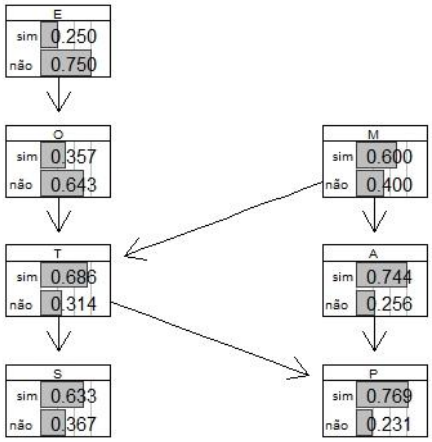


Figura 6.3 – Grafo da rede da Figura 2.9 com as distribuições de cada variável gerado com o programa R.

Nesta secção assumiu-se que conhecíamos a estrutura e os parâmetros da rede em estudo. Contudo, em muitos casos de estudos se parte de uma base de dados inicial, à qual se pretende ajustar uma estrutura em rede bayesiana adequada, através da aplicação de algoritmos de aprendizagem de estrutura e dos parâmetros recorrendo a algoritmos, como os apresentados no capítulo 4.

6.3 Geração de uma base de dados aleatória

Uma vez que não possuimos uma base de dados referente à rede utilizada como exemplo ao longo desta dissertação, de modo a se aplicarem os algoritmos de aprendizagem da estrutura e estimação dos parâmetros, utiliza-se a função `rbn` do *package* `bnlearn`, que permite gerar uma base de dados aleatória para um conjunto de variáveis de um GAD. Deste modo procedemos à simulação de uma base de dados aleatória de dimensão 10 000 da rede

bayesiana apresentada na Figura 6.2 e na Figura 6.3, fixando os dados gerados com o comando `set.seed`:

```
> set.seed(123)

> survey<-rbn(bn, n = 10000, debug = TRUE )
* partial node ordering is: E M O A T S P.
* simulating node E, which doesn't have any parent.
* simulating node M, which doesn't have any parent.
* simulating node O with parents E.
* simulating node A with parents M.
* simulating node T with parents O, M.
* simulating node S with parents T.
* simulating node P with parents T, A.
```

Escrevendo o nome da base de dados obtemos:

```
> survey
     E   O   M   T   S   A   P
1  não sim sim sim sim sim sim
2  sim sim sim não não sim não
3  não não não sim sim sim sim
4  sim sim sim sim sim sim sim
5  sim sim sim sim sim sim sim
6  não não sim sim não sim sim
7  não não não sim sim não sim
 [ reached 'max' / getOption("max.print") -- omitted 9993 rows ]
```

6.4 Estimação de parâmetros

Os parâmetros podem ser estimados com os métodos da máxima verosimilhança ou bayesiano (*cf*. a secção 4.2), respetivamente, pelas linhas de código:

```
> bn.mle<-bn.fit(dag,data = survey,method = "mle")

> bn.bayes<-bn.fit(dag, data = survey, method = "bayes")
```

e as distribuições estimadas de cada variável podem ser visualizadas através dos seguintes comandos:

```
> bn.mle$O
 Parameters of node O (multinomial distribution)

Conditional probability table:

     E
O             sim        não
  sim 0.98275154 0.14910773
  não 0.01724846 0.8508922

> bn.bayes$O
  Parameters of node O (multinomial distribution)

Conditional probability table:

     E
O             sim        não
  sim 0.98265243 0.14913092
  não 0.01734757 0.85086908
```

Note-se que no caso de estimação Bayes podemos incluir um argumento relativo ao `iss`, atribuindo um valor desejado ao mesmo.

6.5 Teste de separação-*d*

O comando dsep permite averiguar num GAD, se duas variáveis são separadas-*d*. Por exemplo, verificamos que as variáveis E e T são independentes dada a variável O, e que as variáveis O e M não são independentes dada a variável T:

```
> dsep(bn,"E","T","O")
[1] TRUE

> dsep(bn,"O","M","T")
[1] FALSE
```

6.6 Testes de independência condicional

Na estrutura guardada no objeto chamado dag as relações de independência condicional definidas pelo critério da separação-*d* são $E \perp T|O$, $S \perp O|T$, $A \perp T|M$, $P \perp M|\{A,T\}$. Considerando uma base dados, os testes de independência condicional permitem verificar se estas relações são suportadas por essa mesma base de dados.

Nestes testes podem ser usados os critérios de informação mútua (`mi`) (*cf.* a expressão (4.6)) e qui-quadrado (`x2`) (*cf.* a expressão (4.7)) e, em ambos os casos, podem ser usadas permutações Monte Carlo normal (`mc-mi/mc-x2`) ou semiparamétrica (`sp-mi/sp-x2`).

O teste de hipóteses H_0: $H_0: E \perp T|O$ vs H_1: $H_1: E \not\perp T|O$ pode ser implementado segundo o critério `mi` através da seguinte linha de código:

```
> ci.test("E", "T","O" , test = "mi", data = survey)
 Mutual Information (disc.)

data:  E ~ T | O
mi = 2.74, df = 2, p-value = 0.2541
alternative hypothesis: true value is greater than 0
```

Note-se que, uma vez não definido no código, o número de graus de liberdade é por defeito igual a 2, e o número de amostras usadas nas permutações Monte Carlo igual a 5000. Os resultados de aplicação dos testes referidos aos conjuntos de variáveis em questão, estão resumidos nas seguintes tabelas:

Tabela 6.1 – Valores-*p* das relações entre as variáveis com os testes `mi`.

Teste	$Valor - p$			*Valor Crítico*
	mi	mc- mi	sp- mi	
$E \perp T\|O$	0.2541	0.2418	0.2049	2.74000
$S \perp O\|T$	0.8278	0.8296	0.8628	0.37791
$A \perp T\|M$	0.4499	0.4514	0.4484	1.59750
$P \perp M\|\{A,T\}$	0.3199	0.3206	0.3330	4.69640

Tabela 6.2 – Valores-*p* das relações entre as variáveis com os testes x2.

$Teste$	$Valor - p$			$Valor\ Crítico$
	x2	mc- x2	sp- x2	
$E \perp T \mid O$	0.2607	0.2636	0.2829	2.6887
$S \perp O \mid T$	0.8330	0.8418	0.8890	0.3655
$A \perp T \mid M$	0.4487	0.4496	0.4398	1.6029
$P \perp M \mid \{A, T\}$	0.3183	0.3348	0.2903	4.7109

Interpretando os resultados dos testes conclui-se que, ao nível de significância 0.05, nenhum valor-*p* é significativo, não se rejeitando as hipóteses nulas de que as independências condicionais referidas sejam válidas.

Em alternativa à realização dos testes referidos anteriores, pode ser utilizada a função `arc.strenght`, que permite realizar automaticamente a medição da "força" de todas as arestas presentes num GAD: se o critério desta medição for um teste de independência condicional o resultado é um valor-*p*; caso seja uma medida de *score*, o resultado é o valor de ganho ou de perda no *score*, resultante da eliminação de cada aresta do GAD. Os comandos de aplicação desta função à nossa estrutura são, usando as medidas de *score* BDE, BIC, AIC (*cf.* a subsubsecção 4.1.2.1):

```
> #Testes de independência que devolvem o valor-p
> arc.strength(dag,data=survey,criterion="x2")
> arc.strength(dag,data=survey,criterion="mi")

> #Testes de score que mostram a alteração no score pela retirada de
cada aresta
> arc.strength(dag,data=survey,criterion="bde")
> arc.strength(dag,data=survey,criterion="bic")
> arc.strength(dag,data=survey,criterion="aic")
```

Os resultados da aplicação destes comandos estão resumidos na seguinte tabela:

Tabela 6.3 – Resultados dos testes da força da relação entre as variáveis.

$Arestas$		$Critério$				
De	$Para$	x2	mi	bde	bic	aic
E	O	0.000000e+00	0.000000e+00	-3084.98256	-3084.5451	-3088.15030
O	T	5.468526e-239	1.046975e-275	-623.07231	-623.9547	-631.16500
M	T	7.056258e-38	3.164238e-37	-73.76184	-74.8334	-82.04374
T	S	0.000000e+00	0.000000e+00	-3665.51750	-3665.3479	-3668.95310
M	A	5.579666e-104	3.295410e-102	-225.39686	-225.7716	-229.37677
A	P	1.853017e-81	2.995216e-74	-159.06497	-160.0839	-167.29428
T	P	2.203775e-205	6.217557e-193	-432.30923	-433.3612	-440.57155

Dos testes de independência condicional qui-quadrado, `x2`, e de informação mútua, `mi`, dados os valores aprendidos, concluem-se que relações entre os vértices ligados pelas arestas são fortes. Dos testes de *score* `bde`, `bic` e `aic` conclui-se que a remoção de qualquer

das arestas resultaria na diminuição das pontuações da rede, provando a força das relações entre as variáveis referidas.

Ao contrário da função anterior, a função `boot.strength` avalia a força das arestas dada a sua frequência empírica sobre um conjunto de estruturas aprendidas a partir de um algoritmo escolhido. Esta calcula a probabilidade de cada aresta, dada a sua direção, apresentada na coluna `strength`, e a probabilidade de cada direção da aresta condicionada à presença da aresta no grafo em ambas as direções, apresentada na coluna `direction`. Para aplicar esta função com o algoritmo PC à rede em questão basta correr o seguinte código:

```
> boot<-boot.strength(survey, algorithm="pc.stable")
```

Os doze primeiros resultados são os seguintes:

```
> boot
   from to strength direction
1     E  O    1.000 0.5500000
2     E  M    0.010 0.5000000
3     E  T    0.220 0.9886364
4     E  S    0.030 0.6666667
5     E  A    0.000 0.0000000
6     E  P    0.000 0.0000000
7     O  E    1.000 0.4500000
8     O  M    0.040 0.2500000
9     O  T    1.000 0.8550000
10    O  S    0.025 0.8000000
11    O  A    0.010 0.5000000
12    O  P    0.365 0.8904110
 [ reached 'max' / getOption("max.print") -- omitted 30 rows ]
```

Uma vez que a lista de resultados é longa, podemos restringir estes, por exemplo, a um conjunto de arestas significativas com força superior a 0.85 e direção superior a 0.5:

```
> #Seleção das arestas mais significativas
> boot[(boot$strength > 0.85) & (boot$direction >= 0.5), ]

   from to strength direction
1     E  O     1.00 0.5500000
9     O  T     1.00 0.8550000
15    M  T     0.98 0.8112245
22    T  S     1.00 0.6625000
24    T  P     1.00 0.7875000
33    A  M     1.00 0.6650000
36    A  P     1.00 0.8875000
```

No resultado anterior as ligações selecionadas não diferem muito das apresentadas na estrutura definida originalmente na Figura 2.9, sendo apenas oposta a orientação da aresta entre as variáveis A e M.

A função `score` permite medir as pontuações com as diferentes medidas apresentadas, sendo que o código `loglik` se refere à verosimilhança logarítmica.

```
> score(dag,data=survey,type="loglik")
[1] -34226.16

> score(dag,data=survey,type="bic")
[1] -34299.85

> score(dag,data=survey,type="aic")
[1] -34242.16

> score(dag,data=survey,type="bde")
[1] -34303.85
```

6.7 Geração de uma estrutura aleatória

O *package* `bnlearn` contém uma função `random.graph`, que gera um grafo aleatório para um conjunto de variáveis. Aplicando essa função às variáveis em questão temos, por exemplo:

```
> dagrdn<-random.graph(nodes =c("E","O","M","T","P","A","S") )
> modelstring(dagrdn)
[1] "[E][O][M][T][P|E][A|E:T][S|O:A]"
```

Se calculamos o *score* de ajustamento deste grafo à base de dados `survey`, constata-se que é inferior, como esperado:

```
> score(dagrdn,data=survey,type="loglik")
[1] -42100.71
```

6.8 Aplicação dos algoritmos de aprendizagem da estrutura

A geração de estruturas aleatórias e a aplicação de testes de independência condicional individual podem ser atividades demoradas. Para contornar este problema são aplicados algoritmos para aprender estruturas que se adequem a uma base de dados.

6.8.1 Algoritmo baseado em restrições: algoritmo PC

De entre os algoritmos baseados em restrições, o *package* `bnlearn` contém o PC, apresentado na subsubsecção 4.1.1.2. A aplicação deste à base de dados gerada acima é concretizada pela seguinte linha de código:

```
> dag.pc<-pc.stable(survey,debug = TRUE,alpha=0.05)
----------------------------------------------------------------
* investigating E - O , d-separating sets of size 0 .
  > neighbours of E : M T S A P
    > node E is dependent on O (p-value: 0).
  > neighbours of O : M T S A P
----------------------------------------------------------------
[…]
----------------------------------------------------------------
* investigating A - P , d-separating sets of size 0 .
  > neighbours of A : E O M T S
    > node A is dependent on P (p-value: 3.17789e-72).
  > neighbours of P : E O M T S
----------------------------------------------------------------
[…]
 [ reached getOption("max.print") -- omitted 27 rows ]
```

Note-se que a opção `debug = TRUE` faz com que todos os passos sejam mostrados. O grafo resultante é um GAPD, como se pode observar quando se executa o comando:

```
> graphviz.plot(dag.pc)
```

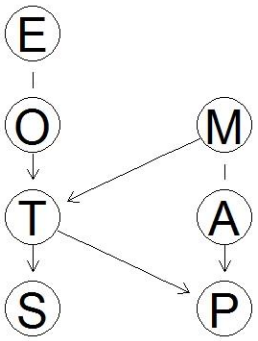


Figura 6.4 – Grafo resultante da aprendizagem com o algoritmo PC usando o programa R.

6.8.2 Algoritmos baseados em funções score: algoritmos HC e tabu

Os algoritmos baseados no *score* presentes no *package* `bnlearn` são os de pesquisa ávida, ou seja, o de subida de colina (`hc`) e tabu (`tabu`), referidos na subsecção 4.1.2. Quando aplicados à base de dados em questão devolvem a estrutura representada na Figura 2.9.

```
> #HC
> dag.hc<-hc(survey,debug = TRUE)
----------------------------------------------------------------
* starting from the following network:

  Random/Generated Bayesian network

  model:
   [E][O][M][T][S][A][P]
  nodes:                                7
  arcs:                                 0
    undirected arcs:                    0
    directed arcs:                      0
  average markov blanket size:          0.00
  average neighbourhood size:           0.00
  average branching factor:             0.00

  generation algorithm:                 Empty

* current score: -42520.73
* whitelisted arcs are:
* blacklisted arcs are:
* caching score delta for arc E -> O (3084.545132).
* caching score delta for arc E -> M (-4.604948).
* caching score delta for arc E -> T (309.307765).
* caching score delta for arc E -> S (205.981501).
* ….
> modelstring(dag.hc)
[1] "[E][M][O|E][A|M][T|O:M][S|T][P|T:A]"

> #TABU
> dag.tabu<-tabu(survey)
> modelstring(dag.tabu)
[1] "[E][M][O|E][A|M][T|O:M][S|T][P|T:A]"
```

6.8.3 Algoritmos híbridos MMHC e variante de SC

O *package* `bnlearn` contém duas funções de algoritmos híbridos, nomeadamente o de subida da colina max-min (`mmhc`) e rsmax2 (`rsmax2`), sendo este último uma variante do algoritmo SC, apresentado na subsecção 4.1.3 com uma única iteração, onde se podem especificar os algoritmos a utilizar em cada passo.

```
> #MMHC
> dag.mmhc<-mmhc(survey)
> modelstring(dag.mmhc)
[1] "[E][M][O|E][A|M][T|O:M][S|T][P|T:A]"

> #RSMAX2
> dag.rsmax2<-rsmax2(survey)
> modelstring(dag.rsmax2)
[1] "[E][M][O|E][A|M][T|O:M][S|T][P|T:A]"
```

Dado que a função `rsmax2` permite selecionar os algoritmos a combinar, podemos selecionar o algoritmo com base em restrições para a parte restritiva do algoritmo, e um algoritmo que selecione a estrutura que maximiza o *score*. Por exemplo, se combinarmos o algoritmo PC com o da subida da colina, obtemos o grafo obtido com os três algoritmos anteriores:

```
> # RSMAX2_PC_HC
> dag.rsmax22<-rsmax2(survey, restrict = "pc.stable",
+                     maximize = "hc", debug = FALSE
> modelstring(dag.rsmax22)
[1] "[E][M][O|E][A|M][T|O:M][S|T][P|T:A]"
```

Note-se que no *package* `bnlearn`, qualquer algoritmo de aprendizagem permite definir na sua função uma `whitelist` e uma `blacklist`, que correspondem respetivamente a listas de arestas orientadas entre certos vértices que serão obrigatoriamente inseridas ou excluídas do GAD final. Deste modo é possível inserir o conhecimento do especialista préviamente à aplicação de algoritmos de aprendizagem da estrutura.

6.9 Inferência pelo algoritmo de árvores junção

Para a realização de inferência o *package* `bnlearn` necessita do apoio de outro chamado `gRain`, cuja única opção para a inferência exata é dada pelo algoritmo de árvores junção, referido na subsecção 3.2.3 (a versão do *package* `gRain` usada é a 1.3-0). Deste modo, depois de chamar ambos os *packages*, e antes da realização de inferência, a rede deve ser convertida numa árvore. Estas operações realizam-se com as seguintes linhas de código:

```
> library(gRain)

> #Construção da árvore de junção
> junction<-compile(as.grain(bn))
> junction
Independence network: Compiled: TRUE Propagated: FALSE
  Nodes: chr [1:7] "E" "O" "M" "T" "S" "A" "P"
```

Ainda antes da realização da inferência exata, podemos usar a função `setEvidence`, que permite introduzir evidência. Para calcular a distribuição de uma variável ou de um conjunto de variáveis, tendo em conta a evidência introduzida na árvore, podemos usar a função `querygrain`.

De seguida são apresentadas as linhas de código que permitem resolver as questões parciais apresentadas na subsecção 3.1.1.

```
> #Resolver CPQ P(S|A=s,T=s)
> jAT<-setEvidence(junction,nodes = c("A","T"),states =
c("sim","sim"))
> dist.probS<-querygrain(jAT,nodes="S",type="joint")
> dist.probS
S
sim não
0.9 0.1

> #Resolver CPQ P(O|S=s,P=s)
> jSP<-setEvidence(junction,nodes = c("S","P"),states =
c("sim","sim"))
> dist.probO<-querygrain(jSP,nodes="O",type="joint")
> dist.probO
O
      sim       não
0.4619996 0.5380004

> #Resolver CPQ P(M|O=s,T=s)
> jAT<-setEvidence(junction,nodes = c("O","S"),states =
c("sim","sim"))
> dist.probM<-querygrain(jAT,nodes="M",type="joint")
> dist.probM
M
      sim       não
0.6326079 0.3673921
```

7 Considerações finais

As redes bayesianas são modelos probabilísticos que através de uma representação gráfica compacta, que suporta as diferentes relações entre as variáveis, permite uma factorização eficiente das distribuições probabilísticas associadas às mesmas, pelo uso da regra de Bayes e da cadeia. Nos capítulos 3, 4 e 5 foram apresentados métodos para realização de inferência, de aprendizagem da estrutura e dos parâmetros, e ainda métodos de análise conflitos, que permitem avaliar o funcionamento dos modelos construídos.

A eficiência das redes bayesianas na resolução de problemas reais levou à sua aplicação a estudos de várias áreas, desde a medicina, à agricultura, passado pela meteorologia, os seguros.

O desenvolvimento de *software* informático veio facilitar a implementação destas redes, facilitando o trabalho dos especialistas nas análises de dados, os processos de aprendizagem das estruturas e dos parâmetros, a realização de inferência e de análises de conflitos. Entre os programas existentes salientam-se os sistemas Hugin e a linguagem R.

Esta dissertação constitui apenas uma possível apresentação deste tema, focando-se em apenas alguns dos possíveis tópicos de análise. Outros assuntos de interessante de análise são, por exemplo, as redes bayesianas gaussianas e mistas (que consideram variáveis com suportes contínuos), a análise de sensibilidade (que permite responder a questões relativas ao impacto de um subconjunto de evidência numa variável hipótese), as técnicas de construção manual (que oferecem ferramentas que facilitam a construção de redes sem recorrer a programas informáticos, permitindo a construção de redes o menos complexas possível) e os diagramas de influência, que são uma extensão das redes bayesianas que permitem a tomada de decisões sob incerteza, incorporando variáveis utilidade e decisão.

Anexos

Anexo A - Exemplo da aplicação do algoritmo *Hugin*

(*Cf*. Subsubsecção 3.2.3.2 Propagação de informação)

De modo a exemplificar a aplicação do algoritmo *Hugin*, consideremos a árvore de junção relativa à rede bayesiana representada pela Figura 2.9.

Uma vez associadas as distribuições de cada vértice aos *clusters*, podemos calcular os potenciais iniciais de cada um:

$$\begin{aligned} \phi_{C_1} &= P(E)P(O|E) = P(E,O), \\ \phi_{C_2} &= P(S|T), \\ \phi_{C_3} &= P(M)P(T|M,O) = P(M,T|O), \\ \phi_{C_4} &= P(A|M), \\ \phi_{C_5} &= P(P|A,T). \end{aligned} \tag{A.1}$$

Igualando o fator inicial de cada separador a 1 temos:

$$\phi_{S_{1,3}} = \phi_{S_{2,3}} = \phi_{S_{3,4}} = \phi_{S_{4,5}} = 1. \tag{A.2}$$

Considere-se a evidência $(M = s, A = n)$ dada pelo vetor *descobertas* $\boldsymbol{e} = (f_1, f_2)$, onde o vetor $f_1 = (1,0)$ é relativo à variável M, com $P(M) = \big(P(M = s), P(M = n)\big)$, e o vetor $f_2 = (0,1)$ é relativo à variável A, com $P(A) = \big(P(A = s), P(A = n)\big)$. Introduzimos a evidência calculando as multiplicações (feitas coordenada a coordenada):

$$\begin{aligned} f_1 \times \phi_{C_3} &= \phi'_{C_3}, \\ f_2 \times \phi_{C_4} &= \phi'_{C_4}, \end{aligned} \tag{A.3}$$

encontrando assim os novos potenciais iniciais dos *clusters*:

$$\begin{aligned} \phi'_{C_3} &= P^*(M,T|O) = P(M,T|O,M = s), \\ \phi'_{C_4} &= P^*(A|M) = P(A|M,A = n). \end{aligned} \tag{A.4}$$

Escolhendo o *cluster* C_5 como raiz, inicia-se a etapa de recolha de informação, passando ordenadamente as mensagens atualizadas dos *clusters* folha até o *cluster* raiz, de C_1 e C_2 para C_3, de C_3 para C_4 , e de C_4 atualizado para C_5, segundo as seguintes operações.

1. Atualizar $\phi_{S_{1,3}}$ e $\phi_{S_{2,3}}$:

$$\phi^*_{S_{1,3}} = \sum_E \phi'_{C_1} = \sum_E P(E,O) = P(O); \tag{A.5}$$

$$\phi^*_{S_{2,3}} = \sum_S \phi_{C_2} = \sum_S P(S|T) = 1. \tag{A.6}$$

2. Atualizar o fator de C_3:

$$\begin{aligned}\phi^*_{C_3} &= \phi'_{C_3}\frac{\phi^*_{S_{1,3}}}{\phi_{S_{1,3}}}\frac{\phi^*_{S_{2,3}}}{\phi_{S_{2,3}}} = \phi'_{C_3}\phi^*_{S_{1,3}}\phi^*_{S_{2,3}} \\ &= P(M,T|O,M=s)P(O) = P(M,O,T|M=s);\end{aligned} \tag{A.7}$$

3. Atualizar $\phi_{S_{3,4}}$:

$$\phi^*_{S_{3,4}} = \sum_O \phi^*_{C_3} = \sum_O P(M,O,T|M=s) = P(M,T|M=s); \tag{A.8}$$

4. Atualizar o fator de C_4:

$$\begin{aligned}\phi^*_{C_4} &= \phi'_{C_4}\frac{\phi^*_{S_{3,4}}}{\phi_{S_{3,4}}} = \phi'_{C_4}\phi^*_{S_{3,4}} \\ &= P(A|M,A=n)P(M,T|M=s) = P(A,M,T|M=s,A=n);\end{aligned} \tag{A.9}$$

5. Atualizar $\phi_{S_{4,5}}$:

$$\phi^*_{S_{4,5}} = \sum_M \phi^*_{C_4} = \sum_M P(A,M,T|M=s,A=n) = P(A,T|M=s,A=n); \tag{A.10}$$

6. Atualizar o fator de C_5:

$$\begin{aligned}\phi^*_{C_5} = \phi^{**}_{C_5} &= \phi_{C_5}\frac{\phi^*_{S_{4,5}}}{\phi_{S_{4,5}}} = \phi_{C_5}\phi^*_{S_{4,5}} \\ &= P(P|A,T)P(A,T|M=s,A=n) = P(A,P,T|M=s,A=n);\end{aligned} \tag{A.11}$$

Note-se que $\phi^*_{C_1} = \phi'_{C_1}$, $\phi^*_{C_2} = \phi'_{C_2}$, , $\phi^*_{S_{1,3}} = \phi'_{S_{1,3}}$, $\phi^*_{S_{1,2}} = \phi'_{S_{1,2}}$.

Depois de realizada a etapa de recolha de informação, passamos à distribuição de informação, passando as mensagens atualizadas desde o *cluster* raiz até os *clusters* folha, segundo as seguintes operações:

1. Atualizar $\phi_{S_{4,5}}$:

$$\phi^{**}_{S_{4,5}} = \sum_P \phi^*_{C_5} = \sum_P P(A,P,T|M=s,A=n) = P(A,T|M=s,A=n); \tag{A.12}$$

2. Atualizar o fator de C_4:

$$\begin{aligned}\phi_{C_4}^{**} &= \phi_{C_4}^{*}\frac{\phi_{S_{4,5}}^{**}}{\phi_{S_{4,5}}^{*}} = P(A,M,T|M=s,A=n)\frac{P(A,T|M=s,A=n)}{P(A,T|M=s,A=n)} \\ &= P(A,M,T|M=s,A=n);\end{aligned} \tag{A.13}$$

3. Atualizar $\phi_{S_{3,4}}$:

$$\phi_{S_{3,4}}^{**} = \sum_{A}\phi_{C_4}^{**} = \sum_{A}P(A,M,T|M=s,A=n) = P(M,T|M=s,A=n); \tag{A.14}$$

4. Atualizar o fator de C_3:

$$\begin{aligned}\phi_{C_3}^{**} &= \phi_{C_3}^{*}\frac{\phi_{S_{3,4}}^{**}}{\phi_{S_{3,4}}^{*}} = P(M,O,T|M=s)\frac{P(M,T|M=s,A=n)}{P(M,T|M=s)} \\ &= P(M,O,T|M=s,A=n);\end{aligned} \tag{A.15}$$

5. Atualizar $\phi_{S_{1,3}}$:

$$\phi_{S_{1,3}}^{**} = \sum_{M,T}\phi_{C_3}^{**} = \sum_{M,T}P(M,O,T|M=s,A=n) = P(O|M=s,A=n); \tag{A.16}$$

6. Atualizar o fator de C_1:

$$\phi_{C_1}^{**} = \phi_{C_1}'\frac{\phi_{S_{1,3}}^{**}}{\phi_{S_{1,3}}^{*}} = P(E,O)\frac{P(O|M=s,A=n)}{P(O)} = P(E,O); \tag{A.17}$$

7. Atualizar $\phi_{S_{2,3}}$:

$$\phi_{S_{2,3}}^{**} = \sum_{M,O}\phi_{C_3}^{**} = \sum_{M,O}P(M,O,T|M=s,A=n) = P(T|M=s,A=n); \tag{A.18}$$

8. Atualizar o fator de C_2:

$$\begin{aligned}\phi_{C_2}^{**} &= \phi_{C_2}^{*}\frac{\phi_{S_{2,3}}^{**}}{\phi_{S_{2,3}}^{*}} = \phi_{C_2}\,\phi_{S_{2,3}}^{**} \\ &= P(S|T)P(T|M=s,A=n) = P(S,T|M=s,A=n).\end{aligned} \tag{A.19}$$

Anexo B – Cálculos auxiliares à análise de conflitos

(*Cf*. Secção 5.1 Análise de conflitos motivada pela evidência)

Depois de chamados os *packages* `bnlearn` e `gRain`, e de definida a rede bayesiana da Figura 2.9, conforme apresentado na subsecção 6.1, para os cálculos da análise de conflitos exemplificada no capítulo 5, foi necessário correr o seguinte código no programa R.

Primeiro calcularam-se as distribuições conjuntas e marginais de cada variável pertencente ao conjunto evidência:

```
> #P(P=s,S=n,M=s)=0.1007128
> dist.probPMS<-querygrain(junction,nodes=c("P","S","M"),type="joint")
> dist.probPMS
, , S = sim

     P
M           sim        não
  sim 0.3446519 0.05516183
  não 0.1923902 0.04087480

, , S = não

     P
M           sim        não
  sim 0.1313783 0.06880792
  não 0.1007128 0.06602220

attr(,"class")
[1] "parray" "array"


> #P(P=s)=0.7691332
> dist.probP<-querygrain(junction,nodes=c("P"),type="joint")
> dist.probP
P
      sim       não
0.7691332 0.2308668


> #P(S=n)=0.3669212
> dist.probS<-querygrain(junction,nodes=c("S"),type="joint")
> dist.probS
S
      sim       não
0.6330788 0.3669212


> #P(M=n)=0.4
> dist.probM<-querygrain(junction,nodes=c("M"),type="joint")
> dist.probM
M
sim não
0.6 0.4
```

Depois aplicou-se a fórmula (5.5) para calcular o conflito da evidência:

```
> confE<-log((0.7691332*0.3669212*0.4)/0.1007128)
> confE
[1] 0.1140924
```

De seguida calcularam-se as distribuições das variáveis evidência, duas a duas, e calcularam-se os conflitos locais e parciais, confirmando-se que a soma iguala a medida de

conflito global, conforme a expressão (5.5). Por exemplo, para as variáveis P e S corremos o seguinte código:

```
> #P(P=s,S=n)=0.2320911
> dist.probPS<-querygrain(junction,nodes=c("P","S"),type="joint")


> dist.probPS
     S
P            sim       não
  sim 0.53704212 0.2320911
  não 0.09603663 0.1348301
attr(,"class")
[1] "parray" "array"

> #conf(Ep,Es)=0.195526
> log((0.7691332*0.3669212)/0.2320911)
[1] 0.195526

> #confE=0.1140924=conf(Ep,Es)+conf({Ep,Es},Em)=0.195526-
0.08143367=0.1140923

> #conf({Ep,Es},Em)=(P(Ep,Es)P(Epm)/P(Es,Em,Ep))=-0.08143367
> log((0.2320911*0.4)/0.1007128)
[1] -0.081433677
```

Note-se que estes cálculos se realizam de forma análoga para os restantes subconjuntos de evidência.

Para a etapa de resolução dos conflitos, usando a variável A como hipótese analisando a existência de casos raros, correu-se o seguinte código:

```
> #P(A=s)=0.744;P(A=n)=0.256
> dist.probA<-querygrain(junction,nodes=c("A"),type="joint")
> dist.probA
A
  sim   não
0.744 0.256

> #P(H|E)=(P(h1|E),P(h2|E))=(0.6994635,0.3005365)
> jPSM<-setEvidence(junction,nodes = c("P","S","M"),states =
c("sim","não","não"))
> dist.probA<-querygrain(jPSM,nodes="A",type="joint")
> dist.probA
A
      sim       não
0.6994635 0.3005365

> #conf(E+h1)=0.1140924+0.06172742=0.1758198
> normloglik1<-log(0.744/0.6994635)
> normloglik1
[1] 0.06172742

> confE+normloglik1
[1] 0.1758198

> #conf(E+h2)=0.1140924-0.1603918=-0.0462994
> normloglik2<-log(0.256/0.3005365)
> normloglik2
[1] -0.1603918

> confE+normloglik2
[1] -0.0462994
```

O código seguinte refere-se à tarefa de medir o impacto da omissão de evidência na variável hipótese A:

```
> #P(H|E\Ep)=(0.63,0.37)
> jMS<-setEvidence(junction,nodes = c("S","M"),states =
c("não","não"))
> dist.probA<-querygrain(jMS,nodes="A",type="joint")
> dist.probA
A
 sim  não
0.63 0.37

> #P(H|E\Es)=(0.6872604,0.3127396)
> jPM<-setEvidence(junction,nodes = c("P","M"),states =
c("sim","não"))
> dist.probA<-querygrain(jPM,nodes="A",type="joint")
> dist.probA
A
      sim       não
0.6872604 0.3127396


> #P(H|E\Em)=(0.7906106,0.2093894)
> jPS<-setEvidence(junction,nodes = c("P","S"),states =
c("sim","não"))
> dist.probA<-querygrain(jPS,nodes="A",type="joint")
> dist.probA
A
      sim       não
0.7906106 0.2093894
```

Os custos de omissão de cada peça de evidência para a variável hipótese A foram calculados através dos seguintes comandos:

```
> #(X=A,Ei=P)=0.01066785
> CAP<-
(0.6994635*(log(0.6994635/0.63)))+(0.3005365*(log(0.3005365/0.37)))
> CAP
[1] 0.01066785

> #(X=A,Ei=S)=0.0003489455
> CAS<-
(0.6994635*(log(0.6994635/0.6872604)))+(0.3005365*(log(0.3005365/0.312
7396)))
> CAS
[1] 0.0003489455

> #(X=A,Ei=M)=0.02292729
> CAM<-
(0.6994635*(log(0.6994635/0.7906106)))+(0.3005365*(log(0.3005365/0.209
3894)))
> CAM
[1] 0.02292729
```

Note-se que estes cálculos se realizam de forma análoga para a variável hipótese T.

Bibliografia

Abramson, B., Brown, J., Edwards, W., Murphy, A., & Winkler, R. L. (1996). Hailfinder: A Bayesian system for forecasting severe weather. *International Journal of Forecasting, 12*, 57-71.

Andreassen, S., Rosenfalck, A., Falck, B., Olesen, K. G., & Andersen, S. K. (1996). Evaluation of the diagnostic performance of the expert EMG assistant MUNIN. *Electroencephalography and Clinical Neurophysiology, 101*, pp. 129-144.

Binder, J., Koller, D., Russell, S., & Kanazawa, K. (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning, 29*, 213-244.

Carvalho, A. M. (2009). *Scoring Functions for Learning Bayesian Networks.* Relatório Técnico, IST, TULisbon/INESC-ID, Lisboa.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. Em D. Fisher, & H.-J. Lenz (Edits.), *Learning from Data, Artificial Intelligence and Statistics V* (pp. 121-130). New York: Springer.

Darwiche. (2008). Bayesian Networks. Em F. van Harmelen, V. Lifschitz, & B. Porter (Edits.), *Handbook of Knowledge Representation* (Vol. 3, pp. 467-509). Elsevier.

Edwards, D. (2000). *Introduction to Graphical Modelling* (2ª ed.). New York: Springer.

Fenton, N., & Neil, M. (2013). *Risk Assessment and Decision Analysis with Bayesian Networks.* Florida: CRC Press.

Friedman, N., Nachman, I., & Pe'er, D. (1999). Learning Bayesian network structure from massive datasets: the «sparse candidate« algorithm. *UAI'99 Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence* (pp. 206-215). Stockholm: Morgan Kaufmann Publishers Inc.

Heckerman, D. (1986). Probabilistic interpretations for MYCIN's certainty factors. *Machine Intelligence and Pattern Recognition, 4*, 167-196.

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

Heckerman, D., Horvitz, E., & Nathwani, B. (1992). Toward normative expert systems: Part I The Pathfinder project. *Methods of Information in Medicine, 31*, 90-105.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks.* London: UCL Press.

Kjærulff, U., & Madsen, A. L. (2008). *Bayesian networks and influence diagrams: A guide to construction and analysis.* New York: Springer.

Kjærulff, U., & Madsen, A. L. (2008). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis.* New York: Springer.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* Cambridge: MIT Press.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability.* (N. Morrison, Ed., & N. Morrison, Trad.) New York: Chelsea Publishing Company. Obtido em 5 de março de 2019, de https://archive.org/details/foundationsofthe00kolm

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2019). *Continuous Multivariate Distributions, Volume 1: Models and Applications* (2ª ed.). New York: Wiley.

Kristensen, K., & Rasmussen, I. A. (2002). The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture, 33*, pp. 197–217.

Le, T. D., Hoang, T., Li, J., Liu, L., Liu, H., & Hu, S. (2016). A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1-13.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence, 29*, 241-288.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* California: Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* New York: Cambridge University Press.

Pearl, J., & Verma, T. S. (1991). A theory of inferred causation. *KR'91 Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning* (pp. 441-452). California: Morgan Kaufmann.

Raiffa , H., & Schlaifer, R. (1961). *Applied Statistical Decision Theory.* Division of Research, Graduate School of Business Administration, Harvard University.

Robinson, R. W. (1977). Counting unlabelled acyclic digraphs. *Lecture Notes in Mathematics: Combinatorial Mathematics V*.

Robinson, R. W. (1977). Counting unlabelled acyclic digraphs. Em C. Little (Ed.), *Combinatorial Mathematics V* (pp. 28-43). Berlin: Springer.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science, 308*, pp. 523-529.

Scutari, M. (2018). Dirichlet Bayesian network scores and the maximum relative entropy principle. *Behaviormetrika, 45*, pp. 337-362.

Scutari, M., & Denis, J.-B. (2014). *Bayesian Networks: With Examples in R.* Florida: CRC Press.

Scutari, M., & Ness, R. (2019). Package 'bnlearn'. *Bayesian network structure learning, parameter learning and inference, R package version 4.4.1*.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction and Search.* Massachusetts: MIT Press.

Tsamardinos, I., & Bourbodakis, G. (2010). Permutation testing improves Bayesian network learning. Em J. Balcázar, F. Bonchi, A. Gionis, & M. Sebag (Edits.), *Machine Learning and Knowledge Discovery in Databases* (pp. 322-337). Berlin: Springer.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning, 65*, 31-78.

Velosa, S., & Pestana, D. (2010). *Introdução à Probabilidade e à Estatística* (4ª ed.). Lisboa: Fundação Calouste Gulbenkian.

Williamson, J. (2002). Probability Logic. Em D. M. Gabbay, R. H. Johnson, H. J. Ohlbach, & J. Woods (Edits.), *Handbook of the Logic of Argument and Inference: The Turn Towards the Practical* (Vol. I, pp. 397 - 424). London: North Holland.

Woods, J., Johnson, R. H., Gabbay, D. M., & Ohlbach, H. J. (2002). Logic and the practical turn. Em D. M. Gabbay, R. H. Johnson, H. J. Olbach, & J. Woods (Edits.), *Handbook of the Logic of Argument and Inference: The Turn Towards the Practical* (Vol. I, pp. 1-39). London: North-Holland.

Índice remissivo