

Modelos Grafos para Expressão Gênica

Cláudia Alexandra Salviano de Medeiros

Modelos Grafos para Expressão Gênica

Cláudia Alexandra Salviano de Medeiros

Orientador: Prof. Dr. Luis Aparecido Milan

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

UFSCar - São Carlos

Julho/2008

**Ficha catalográfica elaborada pelo DePT da
Biblioteca Comunitária da UFSCar**

M488mg

Medeiros, Cláudia Alexandra Salviano de.
Modelos grafos para expressão gênica / Cláudia
Alexandra Salviano de Medeiros. -- São Carlos : UFSCar,
2008.
117 f.

Dissertação (Mestrado) -- Universidade Federal de São
Carlos, 2008.

1. Teoria dos grafos. 2. Expressão gênica. I. Título.

CDD: 511.5 (20^a)

Sumário

Agradecimentos	iv
Resumo	vi
Abstract	vii
1 Introdução	1
2 Grafos	7
2.1 Introdução	7
2.2 Teoria dos Grafos	8
2.2.1 Conceitos Importantes	9
2.2.2 Independência Condicional e Grafos	13
2.2.3 Separação	14
2.2.4 Propriedades de Markov	15
2.3 Modelos Grafos Gaussianos	18
2.3.1 Distribuição Normal Multivariada	18
2.3.2 Função de Verossimilhança	21
2.4 Um Exemplo	24
2.5 Modelos Grafos: Caso Discreto	31
2.5.1 Tabelas de Três Entradas	32

2.5.2	Tabelas Multi-Entradas	37
2.5.3	Grafos e Modelos	40
2.6	Paradoxo de Simpson	40
3	Modelos Grafos Esparsos	45
3.1	Introdução	45
3.2	Grafos, Redes e Regressões	46
3.2.1	Modelos Grafos Gaussianos Esparsos	46
3.2.2	Redes de Dependência	47
3.2.3	Redes Composicionais	47
3.3	Análise e Seleção de Regressões	49
3.3.1	<i>Prioris</i> de Parâmetros de Regressão	49
3.3.2	Verossimilhança Marginal e Distribuições <i>a Posteriori</i>	50
3.4	Construção de Modelos Grafos de Regressões	51
3.4.1	Construção de Modelo	52
3.4.2	Amostragem para Gerar Grafos Múltiplos	54
3.5	Análise de Expressão Gênica em Câncer de Mama	54
3.6	Simulação	56
4	Seleção de Modelo para GAO's via RJMCMC	59
4.1	Introdução	59
4.2	Caso Gaussiano	60
4.2.1	Algoritmo <i>Reversible Jump</i> para o Caso Gaussiano:	61
4.3	Exemplo: Mulher e Matemática	63
4.4	Conclusão	65

5	Computação Estocástica para Modelos Grafos de Alta Dimensão . . .	67
5.1	Introdução	67
5.2	Modelos Grafos e Estrutura Gráfica	69
5.3	Modelos Grafos Gaussianos	72
5.3.1	Fatoração de Densidade e Verossimilhança	72
5.3.2	<i>Prioris</i> e <i>Posterioris</i> para Matrizes de Covariância	73
5.3.3	Funções de Verossimilhança Marginal sobre Grafos	75
5.4	Verossimilhança para Modelos Não-Decomponíveis	76
5.5	Algoritmo de Monte Carlo via Cadeia de Markov	78
5.6	Algoritmo de Busca Estocástica <i>Shotgun</i>	79
5.7	Exemplos Simulados	80
5.7.1	Resultados	82
5.8	Exemplo: Dados de Expressão Gênica	87
5.9	Discussão	89
5.10	Simulação	90
6	Considerações Finais	93
A	Definições	96
B	Linhas de Comandos	98
	Referências Bibliográficas	114

Agradecimentos

À Deus, que me deu paciência, discernimento e sobretudo força para concluir esse trabalho, sem Ele nada teria conseguido.

Aos meus pais que sempre me apoiaram, me incentivaram, não me deixaram desanimar, que mesmo com muitas dificuldades me sustentaram na faculdade e é por isso que hoje estou aqui. Quero agradecer enormemente a Deus por ter colocado vocês na minha vida, vocês são a razão disto tudo.

Aos meus irmãos, que apesar de nossas brigas e discussões, sei que sempre torceram por mim.

Ao Caio, que pacientemente me suportou na reta final deste trabalho, com meus acessos nervosos e crises de choro, que muitas vezes deixou de fazer suas próprias coisas pra me apoiar e me ajudar no desenvolvimento desta dissertação.

Ao meu orientador, Prof. Dr. Luis Aparecido Milan, agradeço o compromisso assumido, as sugestões, os esclarecimentos e comentários sempre oportunos e a compreensão silenciosa dos momentos difíceis pelos quais passei com o andamento deste trabalho, permitindo que meu tempo fluísse, respeitosamente.

Aos meus colegas de mestrado, pelos estudos em grupo, pelas explicações pacientes, pelas conversas descontraídas, pelos churrascos da turma, este trabalho com certeza tem contribuição de todos vocês.

As grandes amigas, Josi, Janaina, Letícia Navarro, Cibele, Juliana, Tatiane, Letícia de Fátima, Taty, Roberta, que estiveram sempre presentes em todos os momentos da minha vida, se não fisicamente, mas em espírito, que sempre torceram por mim, me incentivaram, pela solidariedade, pelo companheirismo nos nossos momentos alegres e tristes, agradeço e sinto-me honrada com a amizade de vocês.

Aos amigos, que sempre me aconselharam e me motivaram.

Ao Fernando Henrique e Nelson, que me ajudaram enormemente na parte computacional deste trabalho.

À Tia Luisa, pelo sorriso e palavras certas nos momentos mais oportunos.

À CAPES, pelo apoio financeiro.

A todos, que de alguma forma, contribuíram para a finalização deste trabalho.

Resumo

Neste trabalho temos como objetivo verificar metodologias estatísticas que podem ser aplicadas a problemas que envolvem um grande número de variáveis, utilizando para isto modelos grafos, que auxiliam na visualização da estrutura de dependência e independência condicional, portanto um modelo grafo representa a relação entre variáveis aleatórias (dependência, independência e independência condicional), onde cada vértice é uma variável aleatória e as arestas entre os vértices representam as diferentes formas de relação entre as variáveis. Nesta dissertação trabalhamos com modelos grafos Gaussianos. Abordamos metodologias para modelos de grande escala (“p grande, n pequeno”) voltadas para análise de associações entre genes, utilizando dados de expressão gênica. Descrevemos os modelos grafos esparsos e implementamos um algoritmo computacional. Verificamos uma abordagem bayesiana usando *Reversible Jump* MCMC. Apresentamos também os modelos grafos decomponíveis quanto a eficiência computacional obtida a partir da decomposição dos espaços amostrais, e encontramos o melhor grafo decomponível baseado no algoritmo de Metropolis-Hastings para um conjunto de dados reais.

Palavras-chave: expressão gênica, modelos grafos esparsos, modelos grafos decomponíveis, *Reversible Jump* MCMC

Abstract

The purpose of this work is to examine statistical methodologies that can be applied to problems that involve a large number of variables using as a tool graphical models that assist on the visualization of the conditional independency and dependency structure, thus a graphical model represents the relationship between random variables (dependence, independence and conditional independence), each node is a random variable and the edges between the nodes are different ways they relate to each other. This dissertation studies Gaussian graphical models. We use methodologies for large scale models (“large p and small n ”) used on the analysis of gene association from gene expression data. We describe the sparse graphical models and we implement a computational algorithm. We verified a Bayesian approach using the Reversible Jump MCMC. We also introduce decomposable graphical models in relation to the computational efficiency attained by the decomposition of the sample space, and we found the best decomposable graph based on the Metropolis-Hastings algorithm for a real data set.

Keywords: gene expression, sparse graphical models, decomposable graphical models, Reversible Jump MCMC

Capítulo 1

Introdução

A Genética é o ramo da biologia que estuda como se transferem as características físicas e biológicas de geração para geração. Muitos cientistas acreditam que a explicação para inúmeros problemas genéticos se encontra nos genes. A hereditariedade é a herança genética que recebemos de nossos antepassados, seja ela, características físicas ou, até mesmo, doenças. Daí a explicação de filhos se parecerem com o pai, com a mãe, avô, avó, tio, tia e até parentes mais distantes.

Gregor Mendel (1822-1884) é considerado o pai da genética. Seus experimentos com *Pisum sativum* (ervilhas) publicados em 1866, realizados no jardim do mosteiro onde trabalhava como professor substituto e as conclusões tiradas na sua investigação constituem o fundamento da genética atual.

Mendel não foi o primeiro a realizar o experimento de hibridação, porém foi o primeiro a considerar os resultados em termos de características individuais. Em seu experimento, Mendel contou e classificou as ervilhas resultantes de cruzamentos, comparou as proporções observadas e formulou hipóteses para explicar as diferenças. Embora Mendel tivesse visualizado um padrão probabilístico para a transmissão das unidades hereditárias, ele não conceituou o mecanismo biológico envolvido. Com base nos seus experimentos e hipóteses, ele predisse e, verificou suas predições com os resultados de cruzamentos posteriores.

De acordo com Pazza (2006), a genética é a ciência que estuda a estrutura e funcionamento dos ácidos nucleicos, do DNA (ácido desoxirribonucleico) e do RNA (ácido

ribonucléico). Estuda os processos de divisão celular, onde o DNA é replicado (produz uma réplica de si mesmo) para ser repassado às células filhas, os mecanismos de síntese de proteínas (transcrição e tradução), bem como as leis da hereditariedade e a dinâmica dos genes nas populações.

O que facilita ligeiramente o estudo genético dos organismos vivos é a universalidade do DNA, ou seja, dos processos de manutenção e expressão das características de hereditariedade, os quais são compartilhados entre os mais diferentes organismos vivos, salvo algumas pequenas particularidades, principalmente em se tratando de procariotos (bactérias) x eucariotos (fungos, animais, plantas).

Na maioria dos organismos eucariotos, o DNA é formado por duas fitas complementares de nucleotídeos em forma de espiral, chamada dupla hélice. A dupla hélice está organizada com proteínas no núcleo da célula eucariótica, formando um cromossomo. O número de cromossomos ¹ (moléculas de DNA enroladas com proteínas) pode variar de organismo para organismo. Os seres humanos, por exemplo, tem 46 cromossomos ou 46 moléculas de DNA enrolado com proteínas. Organismos chamados diplóides são aqueles que recebem um lote cromossômico do pai e um da mãe, e representam uma grande parcela dos organismos vivos atuais.

A estrutura espacial do DNA foi descoberta em 1953 por James Watson e Francis Crick, através de estudos de difração de raios-X. O DNA é formado por nucleotídeos, que se compõem de açúcar, fosfato e uma base nitrogenada. Existem quatro bases nitrogenadas no DNA, as quais se unem aos pares, são elas: adenina (A), timina (T), guanina (G) e citosina (C), em que A e T se atraem mutuamente, da mesma forma C e G. Elas obedecem rigorosamente essa regra de que só podem se unir destas duas maneiras: A se liga a T e G se liga a C. Não pode existir no DNA um par de bases formado de adenina e citosina, ou de timina e guanina, por exemplo. A ordem particular em que as bases se alinham ao longo da cadeia de açúcar e fosfato é chamada de seqüência nucleotídica do DNA. Essa seqüência é característica para cada organismo e codifica milhões de sinais que a célula consegue interpretar como instruções para a fabricação de proteínas.

Nos humanos, cada indivíduo recebe 23 cromossomos do pai e 23 cromossomos

¹Apresentação organizada compactada do DNA. Proteínas e RNA estão envolvidos no enrolamento da fita de DNA, e formam juntos os cromossomos.

da mãe, isto é, os humanos têm 23 pares de cromossomos. Cada cromossomo que é proveniente do pai tem um cromossomo correspondente proveniente da mãe, chamados homólogos, formando o par de cromossomos, exceto para os cromossomos sexuais X e Y, pois as mulheres têm dois cromossomos X enquanto os homens têm um cromossomo X e um Y. Os cromossomos homólogos pareiam-se na meiose (processo de divisão celular que gera gametas com apenas um lote cromossômico) e trocam trechos de seus DNA's em um processo chamado de recombinação (ou *crossing-over*).

As seqüências de nucleotídeos, em uma molécula de DNA dos cromossomos, são responsáveis por características específicas por exemplo a cor do olho, são chamadas de genes². Um determinado gene ocupa portanto, uma região específica de um cromossomo, o que chamamos de *locus*. Um mesmo cromossomo pode ter centenas de *loci*. Como organismos diplóides recebem um lote cromossômico do pai e outro da mãe, em ambos os cromossomos do par de homólogos, temos a mesma região para um determinado gene, e estas seqüências podem ser iguais ou diferentes.

Cada seqüência nestes *loci* é chamada de alelo, ou seja, alelos são genes localizados em um determinado *locus* de um dos cromossomos do par de homólogos. Assim, para uma dada característica determinada por um único gene, um indivíduo diplóide recebe um alelo do pai e outro da mãe, podendo ter no máximo dois alelos para esta característica (ou para cada *locus* gênico). Quando os dois alelos são iguais, dizemos que o indivíduo é homozigoto para a característica. Quando são diferentes, dizemos que o indivíduo é heterozigoto. O conjunto dos alelos de uma determinada característica é chamado de genótipo. A combinação deste genótipo determina a característica ou fenótipo.

O corpo humano conta com 20 aminoácidos diferentes, que se unem em diferentes seqüências, para construir as diferentes proteínas necessárias à sua entrada e funcionamento. O organismo humano pode sintetizar milhares de diferentes proteínas.

A instrução para que as células fabriquem uma proteína específica é dada por um segmento da cadeia de DNA (gene) que contem uma seqüência específica de bases. Isto é, o que constitui o DNA que contem a mensagem completa para a síntese de uma proteína, em que um gene funciona como uma “sentença”, cujas as letras seriam as quatro bases A, C, G e T. Cada conjunto de três bases (*codons*), na seqüência ao longo do DNA, indicam

²Segmento de DNA que contem informações para produção de proteínas.

às células que um determinado aminoácido deve ser usado na síntese da proteína. Por exemplo, a seqüência de bases **ATG** codifica o aminoácido metionina, e **GAGATGGCA** codifica uma seqüência de três aminoácidos, que são, respectivamente, ácido glutâmico, metionina e alanina.

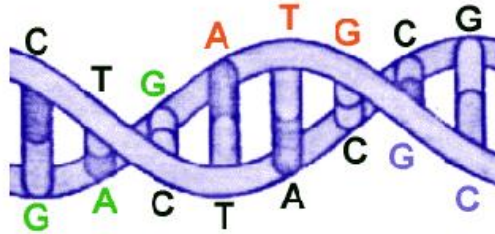


FIGURA 1.1: Fragmento de DNA.

Fonte: <http://www.geocities.com/~sabio/genoma/projetogenoma2.htm>

Desvendar o seqüenciamento das bases do DNA, para cada organismo, é desvendar o seu código genético, o “segredo” de sua formação e de seu funcionamento, pois o DNA é o “manual de instruções” usado pela célula.

Estimava-se que o padrão genético da espécie humana - o genoma humano - contivesse de 60 a 100 mil genes, cada um deles contendo instruções sobre como as células devem produzir um determinado tipo de proteína. Entretanto, em fevereiro de 2001, duas equipes independentes anunciaram simultaneamente a transcrição quase completa do código genético humano e o número de genes calculado foi revelado bem menor: os pesquisadores do Projeto Genoma Humano - projeto desenvolvido pelo consórcio de instituições públicas - anunciaram a existência de cerca de 31 mil genes, e os pesquisadores da Celera - empresa privada americana - anunciaram a existência de cerca de 39 mil genes.

Além da codificação genética, tem havido grandes progressos na descoberta de genes associados a doenças. Supõe-se que as 20 doenças mais comuns, que matam cerca de 80% da população, estejam associadas com aproximadamente 200 genes que compõe o corpo humano. A iniciativa privada tem se dedicado mais intensamente ao estudo desses genes específicos e as indústrias farmacêuticas, especialmente, disputam esse conhecimento que deverá levar a um grande aperfeiçoamento da medicina.

Neste trabalho, temos por objetivo verificar metodologia estatística que possa ser

utilizada, entre outras finalidades, para identificar as relações entre genes, usando modelos grafos. O intuito básico deste trabalho consiste em através dos modelos grafos, providenciar a visualização das dependências e independências condicionais das variáveis.

No Capítulo 2, baseados em Whittaker (1990) e Edwards (1995) fazemos uma revisão da teoria dos grafos. Apresentamos uma introdução aos Modelos Grafos probabilísticos, que são utilizados para visualizar a estrutura de dependência e independência condicional existente entre as variáveis de um vetor aleatório, através de arestas e vértices. Descrevemos a seguir algumas propriedades markovianas dos grafos.

Consideramos os Modelos Grafos Gaussianos e os seus estimadores de máxima verossimilhança e apresentamos um exemplo de aplicação da metodologia a um conjunto de dados, mostrando a relação entre as notas de alunos em cinco disciplinas. Descrevemos modelos grafos para tabelas de contingência multi-entradas, baseadas em amostragem da distribuição Multinomial, e um exemplo onde é possível identificar o paradoxo de Simpson.

Com base no trabalho de Dobra, Hans, Jones, Nevins e West (2003), no Capítulo 3 discutimos a estrutura e metodologia para modelos grafos de grande escala, motivados pela modelagem e exploração de associações entre genes, utilizando dados de expressão gênica.

Na análise da expressão gênica temos um grande número de variáveis e poucas observações para cada variável (“p grande, n pequeno”) e esta situação desafia a ciência estatística moderna, tanto conceitualmente como computacionalmente.

A noção de esparsidade de modelos grafos para dados de expressão gênica reflete uma perspectiva na qual grafos para um gene G , são bem preditos por um subconjunto relativamente pequeno de genes. Temos um exemplo de Dobra *et al.* (2003) que, para o câncer de mama, relaciona o receptor estrógeno e a proliferação e controle da célula Rb - E2F. Terminamos com um estudo de simulação, onde aplicamos o algoritmo de Dobra, Hans, Jones, Nevins e West (2003) para o conjunto de dados das notas de 88 estudantes em 5 disciplinas.

No Capítulo 4 revisamos uma abordagem bayesiana usando *Reversible Jump* MCMC (RJMCMC) de Fronk (2002), introduzido por Green (1995), com o objetivo de ampliar nosso conhecimento e verificar mais uma metodologia que pode ser aplicada para casos em

que é considerada uma variação de espaço muito grande. Esta abordagem possibilita combinar a seleção de modelos com a estimação de parâmetros. Além disso, esta metodologia pode ser estendida para problemas onde ocorrem, dados *missing* ou variáveis ocultas. (ver Fronk, 2002).

No Capítulo 5, discutimos a estrutura e o contexto de modelos grafos não - orientados, modelos Gaussianos, modelos decomponíveis, além de alguns exemplos utilizando vários métodos de computação estocástica. Esta discussão representa uma revisão da estrutura destes modelos grafos e detalhes da existência de métodos computacionais, tais como os métodos MCMC. Segundo Jones, Carvalho, Dobra, Hans, Carter e West (2004) há necessidade para novos métodos computacionais em dimensões maiores e desenvolvimentos alternativos para computação e modelagem gráfica. Apresentamos uma estudo de simulação onde relacionamos 5963 variáveis, utilizando o algoritmo metropolis-hastings.

Capítulo 2

Grafos

2.1 Introdução

Ao contrário de muitos ramos da matemática, nascidos de especulações puramente teóricas, a teoria dos grafos tem sua origem no confronto de problemas práticos relacionados a diversas especialidades e na construção de propriedades comuns, dentre os conceitos relacionados a esses problemas. Basicamente a teoria trata das relações existentes entre os elementos de um ou mais conjuntos. Sua abrangência é muito ampla e muitos problemas envolvendo inter-relacionamento de elementos, seja na química orgânica, eletricidade, organização, transporte, psico-sociologia, genética, etc, podem ser estudados com o auxílio de grafos - devido a existência de associações ou de correspondência entre os elementos do problema (átomos em uma molécula, componentes em um circuito, funcionários em uma empresa, estradas unindo cidades, crianças em uma turma de escola, genes, etc).

O primeiro problema cuja solução envolveu conceitos do que veio a ser a teoria dos grafos foi resolvido por Euler em 1736 e não passava de uma especulação matemática. O chamado “problema das pontes de Königsberg” era análogo aos atuais quebra-cabeças para crianças, baseados em um desenho cujas linhas devem ser percorridas sem que se tire o lápis do papel e sem passar duas vezes sobre a mesma linha. No problema estudado por Euler (Figura 2.1), os pontos correspondiam às margens de um rio e a ilhas, e as linhas as pontes. Ele mostrou que dada a disposição existente das pontes, era impossível percorrer todas passando uma só vez em cada ponte.

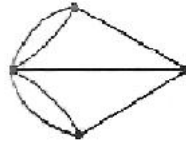


FIGURA 2.1: Pontes de Königsberg.

A modelagem por grafos é uma forma de análise multivariada que utiliza grafos para representar modelos. Sua origem pode derivar da análise de caminhos (Wright, 1921), que tem sido usada para o desenvolvimento atual dos artigos de Darroch, Lauritzen e Speed (1980) e de Lauritzen e Wermuth (1989). As características principais dos modelos grafos são:

- *Interpretação*: os grafos apontam a informação relacionada com a estrutura de independência condicional existente entre as variáveis objetos de estudo. Tanto a dependência quanto a independência condicional são as chaves teóricas dos modelos grafos, combinadas com as propriedades de Markov, que determinam um conjunto de regras explícitas para interpretar os grafos de independência.
- *Simplificação*: qualquer procedimento sistemático para analisar observações multivariadas deverá condensar o conjunto de dados sem eliminar ou obscurecer as associações relevantes existentes. Os modelos grafos englobam o conjunto de todos os possíveis modelos de interação, sendo estimados e interpretados, reduzindo notavelmente a complexidade no processo de seleção dos modelos.

2.2 Teoria dos Grafos

Um grafo $G(K,E)$ é um objeto matemático que consiste de dois conjuntos, um conjunto de vértices, K , e um conjunto E finito de arestas (i, j) , que representam ligações entre vértices de K . Usualmente, tomamos K como sendo um conjunto de números naturais $\{1, 2, \dots, k\}$. Assim, há uma aresta orientada ou seta entre vértices $i, j \in K$ se o conjunto E contém o par ordenado (i, j) . Há uma aresta não orientada ou linha entre estes vértices se E contém ambos os pares (i, j) e (j, i) . O grafo é não orientado se todas as arestas são não orientadas.

Consideraremos somente grafos acíclicos, aqueles sem *loops* ou arestas múltiplas, quando existe mais de uma aresta entre o mesmo par de vértices. O diagrama do grafo é uma figura, na qual círculos ou pontos representam vértices, uma linha representa uma aresta não orientada, e uma seta representa uma aresta orientada.

Exemplo 2.1. O grafo $G(K,E)$ com $K = \{1, 2, 3, 4\}$ e $E = \{(1, 2), (2, 1), (1, 3), (4, 3)\}$ tem o seguinte diagrama mostrado na Figura 2.2



FIGURA 2.2: Grafo Orientado.

2.2.1 Conceitos Importantes

- *Adjacência:* os vértices i e j , são adjacentes se existe uma aresta não orientada entre i e j . Assim, no Exemplo 2.1, os vértices 1 e 2 são adjacentes, mas os vértices 3 e 4, e 1 e 3 não são adjacentes. O grafo não orientado G^u associados com G é o grafo obtido da substituição de todas as setas em G por linhas.
- *Subgrafo:* um grafo $H(K', E')$ é um subgrafo de um grafo $G(K, E)$ se todos os vértices e todas as arestas de H pertencem a $G(K' \subseteq K, E' \subseteq E)$, e cada aresta de H possui as mesmas extremidades que em G . Denotamos um subgrafo utilizando a notação usada para conjuntos, $H \subset G$. Considere o grafo da Figura 2.3, alguns subgrafos são apresentados na Figura 2.4

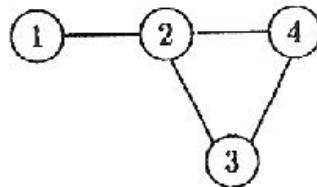


FIGURA 2.3: Grafo não orientado com 4 vértices.

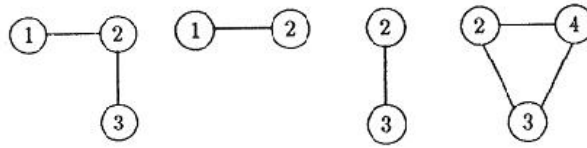


FIGURA 2.4: Subgrafos da Figura 2.3.

Observações:

1. Todo grafo é um subgrafo de si mesmo;
 2. Um subgrafo de um subgrafo de G também é um subgrafo de G ;
 3. Um vértice de G é um subgrafo de G ;
 4. Uma aresta, juntamente com seus vértices é um subgrafo.
- *Grafo Completo:* é um grafo acíclico tal que existe uma aresta para cada par de vértices. Se G possui n vértices, denotamos o grafo completo por K_n . Alguns exemplos de grafos completos são apresentados na Figura 2.5:

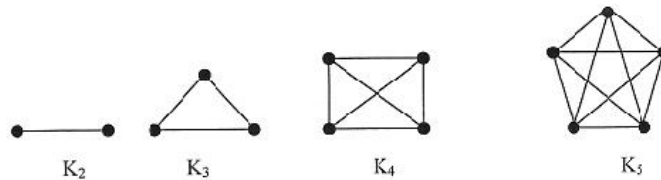


FIGURA 2.5: Grafos completos.

- *Clique:* é um subgrafo completo de G . Considerando, o grafo da Figura 2.6

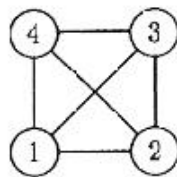


FIGURA 2.6: Grafo completo com 4 vértices.

alguns cliques são mostrados na Figura 2.7

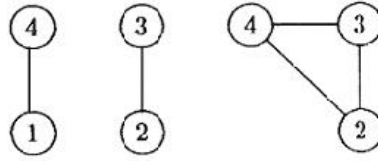


FIGURA 2.7: Alguns cliques da Figura 2.6.

- *Trajeto*: Dado um grafo $G(K, E)$, um trajeto em G consiste de uma seqüência finita alternada de vértices e arestas, começando e terminando por vértices, tal que cada aresta aparece apenas uma vez. Considerando a Figura 2.8

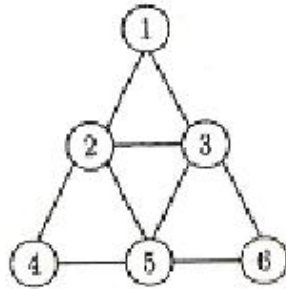


FIGURA 2.8: Grafo com seis vértices.

um trajeto pode ser observado pela linha tracejada, que corresponde ao conjunto $\{(1, 2); (2, 4); (4, 5); (5, 2); (2, 3); (3, 5); (5, 6)\}$:

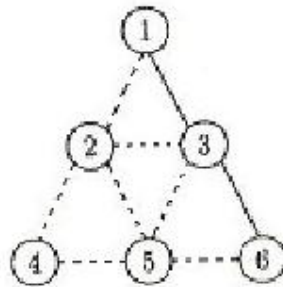


FIGURA 2.9: Trajeto contido na Figura 2.8 representado pela linha tracejada.

- *Caminho*: Dado um grafo $G(K, E)$, um caminho em G consiste de uma seqüência finita alternada de vértices e arestas, começando e terminando por vértices, tal que cada aresta é incidente ao vértice que a precede e ao que a sucede, e não há repetição de vértices. Em outras palavras, um caminho é um trajeto onde não há repetição de vértices. Considerando o grafo da Figura 2.8, vemos que um caminho deste grafo pode ser observado pela linha tracejada, que corresponde ao conjunto $\{(1, 2); (2, 4); (4, 5); (5, 3)\}$:

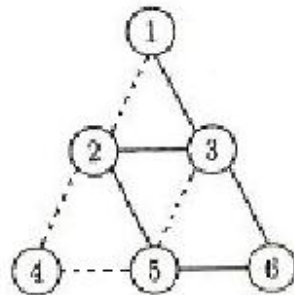


FIGURA 2.10: Caminho contido na Figura 2.8 representada pela linha tracejada.

- *Trajeto Fechado*: É um trajeto no qual o vértice inicial e o final são iguais. Considerando o grafo da Figura 2.8, vemos que um trajeto fechado deste grafo pode ser observado pela linha tracejada, que corresponde ao conjunto $\{(1, 2); (2, 4); (4, 5); (5, 2); (2, 3); (3, 1)\}$:

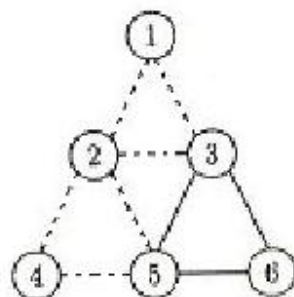


FIGURA 2.11: Trajeto Fechado contido na Figura 2.8 representado pela linha tracejada.

- *Ciclo ou Circuito*: É um trajeto fechado no qual nenhum vértice, com exceção do inicial e do final, aparece mais de uma vez. Considerando o grafo da Figura 2.8,

vemos que um circuito deste grafo é dado pela linha tracejada, que corresponde ao conjunto $\{(1, 2); (2, 4); (4, 5); (5, 6); (6, 3); (3, 1)\}$:

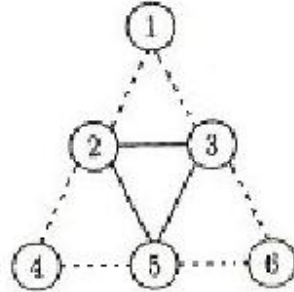


FIGURA 2.12: Circuito contido na Figura 2.8 representado pela linha tracejada.

Observação 2.1. : Grafos que contém ciclos são chamados de *grafos cíclicos*, e grafos que não contém ciclos, são conhecidos como *grafos acíclicos*.

2.2.2 Independência Condicional e Grafos

Grafos são instrumentos adequados para representar as relações de independência condicional em variáveis. Seja $X = (X_1, X_2, \dots, X_k)$ um vetor de variáveis aleatórias e $K = (1, 2, \dots, k)$ o conjunto de vértices correspondentes. Temos um grafo independente, ou mais precisamente um grafo condicionalmente independente se não há aresta entre variáveis sempre que o par de variáveis seja condicionalmente independente dado todas as demais variáveis. O vetor das variáveis restantes é chamado de resto.

Usamos a notação $1 \perp\!\!\!\perp 2 \mid \{3, 4\}$ para $X_1 \perp\!\!\!\perp X_2 \mid (X_3, X_4)$, ou X_1 é condicionalmente independente de X_2 dado X_3 e X_4 , tal que a independência condicional de X_i e X_j dado o resto pode ser escrita como $i \perp\!\!\!\perp j \mid K \setminus \{i, j\}$; ocasionalmente escrevemos $X_i \perp\!\!\!\perp X_j \mid \text{resto}$.

Definição 2.1. : Um grafo que represente as relações de independência condicional de X é um grafo não orientado $G = (K, E)$ onde $K = \{1, 2, \dots, k\}$ e (i, j) não está no conjunto de arestas E se, e somente se, $X_i \perp\!\!\!\perp X_j \mid X_K \setminus \{i, j\}$.

Exemplo 2.2. : Tome $k = 4$, e considere a função densidade de $X = (X_1, X_2, X_3, X_4)$, como sendo $f_X(x) = \exp(u + x_1 + x_1x_2 + x_2x_3x_4)$, $x; x = (x_1, x_2, x_3, x_4)$, $0 < x_i < 1$,

$i = 1, 2, 3, 4$, onde a constante u garante a densidade integrada para 1. A aplicação direta do critério da fatoração¹ implica que $f_X(x)$ pode ser fatorada como

$$f_X(x) = \exp(u + x_1 + x_1x_2)\exp(x_2x_3x_4)$$

onde $g(x_1, x_2) = \exp(u + x_1 + x_1x_2)$ e $h(x_2, x_3, x_4) = \exp(x_2x_3x_4)$, o que nos garante

$$X_1 \perp\!\!\!\perp X_4 | (X_2, X_3) \text{ e } X_1 \perp\!\!\!\perp X_3 | (X_2, X_4),$$

e conseqüentemente o grafo é dado pela Figura 2.13

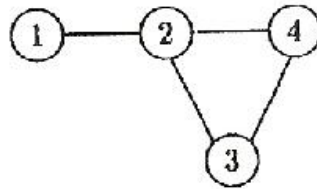


FIGURA 2.13: Grafo associado ao Exemplo 2.2.

2.2.3 Separação

Um ponto importante na independência condicional de grafos certamente é a interdependência das variáveis, em particular, quando duas variáveis são adjacentes ou não, e se são, como elas são separadas. Isto nos leva a concluir que variáveis adjacentes são independentes dado a separação de algumas delas. O teorema da separação providencia uma justificação para esta interpretação.

Exemplo 2.3. : Considere a independência condicional que pode ser vista na Figura 2.14



FIGURA 2.14: Grafo não orientado.

Uma maneira na qual o grafo condicionalmente independente poderia ser definido é

$$X_1 \perp\!\!\!\perp X_3 | (X_2, X_4),$$

¹Critério da fatoração segue no Apêndice A

mas nós gostaríamos de concluir que

$$X_1 \perp\!\!\!\perp X_3 | X_2$$

onde X_4 seria omitido do condicionamento porque o grafo sugere que esta variável é redundante na explicação da dependência entre X_1 e X_3 . Esta condição é dada pelo teorema da separação.

Teorema 2.1. (*Teorema da Separação*): Se X_a , X_b e X_c são vetores que contém subconjuntos disjuntos de variáveis de X , e se, no grafo independente de X , cada vértice em b é separado de cada vértice em c por um subconjunto a , então

$$X_b \perp\!\!\!\perp X_c | X_a.$$

A recíproca do teorema da separação é imediata: se é certo que para qualquer a que separe i e j se cumpra que $i \perp\!\!\!\perp j | a$, então será certo que para qualquer i e j não adjacentes, $i \perp\!\!\!\perp j | \text{resto}$. De qualquer modo, este resultado é trivial porque o “resto” será sempre o conjunto separador para os vértices não adjacentes.

2.2.4 Propriedades de Markov

- **Propriedade de Markov para Pares de Variáveis:** pares de variáveis não adjacentes no grafo são independentes dado o resto das variáveis. Ou seja, $X_i \perp\!\!\!\perp X_j | K \setminus \{i, j\}$.

Considerando a Figura 2.15

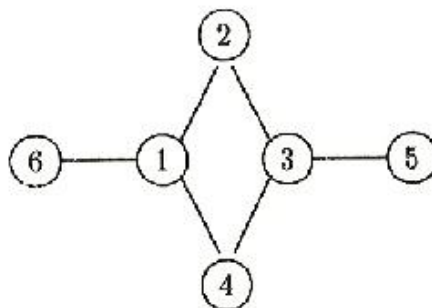


FIGURA 2.15: Grafo com seis vértices.

temos que, $X_1 \perp\!\!\!\perp X_3 | \{X_6, X_2, X_4, X_5\}$ ou $X_4 \perp\!\!\!\perp X_5 | \{X_6, X_2, X_1, X_3\}$.

- **Propriedade de Markov Local:** condicionando somente sobre as variáveis adjacentes, qualquer variável é independente de todas as restantes. Ou seja, para cada vértice i , sendo a o conjunto de vértices adjacentes a i e sendo b o conjunto de vértices restantes, temos $X_i \perp\!\!\!\perp X_b | X_a$ onde $b = K \setminus (\{i\} \cup a)$. Considerando o grafo G da Figura 2.15, temos

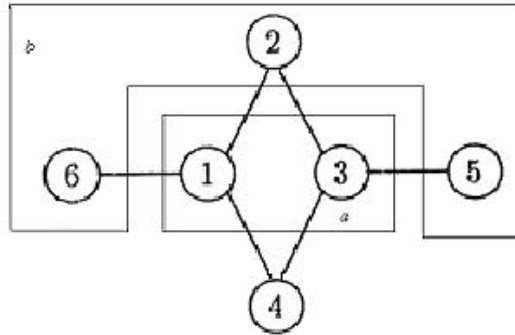


FIGURA 2.16: Grafo com Propriedade de Markov Local.

onde $a = \{X_1, X_3\}$ e $b = \{X_6, X_2, X_5\}$, desta forma $X_4 \perp\!\!\!\perp X_b | X_a$.

- **Propriedade de Markov Global:** subconjuntos de variáveis quaisquer, separados por uma terceira variável, são condicionalmente independentes somente sobre esta terceira variável. Ou seja, se b e c são subconjuntos de vértices separados pelo subconjunto a , então $X_b \perp\!\!\!\perp X_c | X_a$. Considerando o mesmo grafo G ,

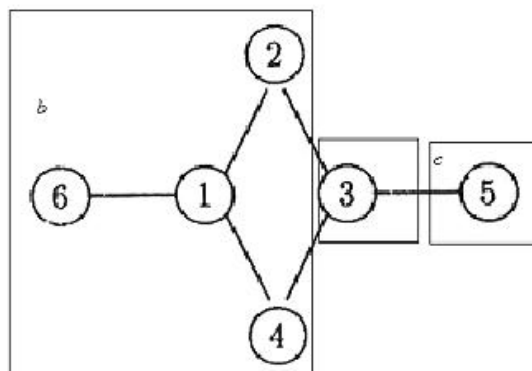


FIGURA 2.17: Grafo com Propriedade de Markov Global.

onde $b = \{X_6, X_2, X_1, X_4\}$ e $c = \{X_5\}$, portanto $X_b \perp\!\!\!\perp X_c | X_3$.

Propriedades de Markov Equivalentes

De acordo com Whittaker (1990), temos o seguinte teorema

Teorema 2.2. As propriedades de Markov para pares de variáveis, de Markov local e de Markov global, são equivalentes.

Demonstração. Para resumir a prova dessas equivalências, nós argumentamos:

1. A propriedade de Markov global implica a propriedade de Markov local, porque o conjunto limitante é sempre um conjunto de separação.
2. A propriedade de Markov local implica a propriedade de Markov para pares de variáveis, porque se o grafo com vértices $K = \{1, 2, \dots, k\}$ satisfaz a propriedade de Markov local, então para cada vértice i , com conjunto limitante a ,

$$X_i \perp\!\!\!\perp X_b | X_a$$

onde b denota os vértices restantes, isto é, $b = K \setminus (\{i\} \cup a)$. Seleccionando um vértice j não adjacente a i , desta forma j estará em b , colocando $c = b \setminus \{j\} = K \setminus (\{i, j\} \cup a)$ e reescrevendo a independência assim,

$$X_i \perp\!\!\!\perp (X_j, X_c) | X_a.$$

Pelo lema do bloco independente, isto é equivalente a

$$X_i \perp\!\!\!\perp X_j | (X_a, X_c) \text{ e } X_i \perp\!\!\!\perp X_c | (X_a, X_j).$$

Mas $a \cup c = K \setminus \{i, j\}$ é o resto, e a primeira independência é exatamente a propriedade de independência para pares de variáveis.

3. O teorema da separação afirma que a propriedade de independência condicional para pares de variáveis implica a propriedade de Markov global.

□

2.3 Modelos Grafos Gaussianos

Considere um vetor aleatório k - dimensional $X = (X_1, X_2, \dots, X_k)$ e um grafo $G = (K, E)$. Um modelo grafo para X é uma família de distribuições de probabilidade para X , que satisfaz as restrições de independência condicional para pares de variáveis em G . Quando a distribuição é normal multivariada então falamos do modelo grafo Gaussiano. Neste caso, temos que independências condicionais são equivalentes à especificação de zeros na matriz de covariâncias inversa.

Um procedimento segundo Whittaker (1990) para modelagem com grafos é

1. Estimar a matriz de covariâncias $V = var(X)$ pela matriz de covariâncias amostral, S .
2. Calcular a inversa S^{-1} , a inversa da matriz de correlação amostral. Os elementos da diagonal são interpretados em termos de variâncias parciais: $1/var_N(X_i|resto)$ ou como $var_N(X_i)/var_N(X_i|resto)$ se a matriz de correlação amostral é usada, e indicar como a variável X_i pode ser predita pelas outras variáveis.
3. Escalonar S^{-1} até obter uma diagonal unitária e calcular as correlações parciais amostrais, $corr_N(X_i, X_j|resto)$.
4. Tomar como zero qualquer elemento suficientemente pequeno da inversa escalonada. Desenhar o grafo de independência resultante, conforme a regra que, uma aresta não é incluída no grafo quando o coeficiente de correlação parcial é zero.

2.3.1 Distribuição Normal Multivariada

Definição 2.2. : O vetor k -dimensional X tem distribuição Normal multivariada com função densidade da forma:

$$f_X(\mathbf{x}) = (2\pi)^{-\frac{k}{2}} \det(D)^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(x - \mu)^T D (x - \mu)] \right\}, \text{ para todo } \mathbf{x} \in \mathfrak{R}^k$$

onde μ é o vetor de médias e V é a matriz de covariâncias, com $D = V^{-1}$. Escrevemos que $X \sim N_k(\mu, V)$ para denotar que X tem distribuição normal multivariada.

Observação 2.2. De acordo com Whittaker (1990), dada uma matriz positiva definida simétrica, podemos sempre encontrar a matriz positiva definida A que pode ser expressa como AA^T por decomposição de Clolesky. Como D^{-1} é simétrica e positiva definida, podemos então encontrar A tal que $D^{-1} = AA^T$.

Proposição 2.1. : Algum vetor aleatório $X \sim N(\mu, V)$ pode ser obtido do vetor aleatório Z com Normal multivariada padrão, $Z \sim N(0, I)$, utilizando a transformação: $X = AZ + \mu$.

Corolário 2.1. : O valor esperado e a variância de X são $E(X) = \mu$, e $var(X) = D^{-1}$.

Demonstração. : Se $X = AZ + \mu$, pela linearidade da esperança $E(X) = AE(Z) + \mu$. Mas se $E(Z) = 0$ e $var(Z) = I$, segue que $var(X) = var(AZ) = Avar(Z)A^T = AA^T$ e, por construção, $D^{-1} = V$. \square

O parâmetro da variância $D^{-1} = V$ é a matriz $k \times k$ de variâncias e covariâncias dado por $V = var(X)$. A notação de particionamento estende os parâmetros da variância: $V_{aa} = var(X_a)$, $V_{ab} = cov(X_a, X_b)$, e mais geralmente,

$$V_{a \cup b, a \cup b} = \begin{pmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{pmatrix}$$

Denotamos o valor de $V_{\{i\}, \{j\}}$ por v_{ij} e $v_{ij} = cov(X_i, X_j)$. O elemento correspondente de D é d_{ij} .

Condições simples para independência são derivadas em termos da covariância e da inversa da variância.

Corolário 2.2. (*Independência e covariância zero*): Os vetores aleatórios normais X_a e X_b são independentes, ou $X_a \perp\!\!\!\perp X_b$ se, e somente se,

1. $cov(X_a, X_b) = 0$, ou parametricamente expressado, $V_{ab} = 0$; ou
2. $D_{ab} = 0$, onde D é a inversa variância.

Corolário 2.3. (*Independência condicional e covariância parcial zero*). Os vetores normais X_b e X_c são condicionalmente independentes em X_a , ou $X_b \perp\!\!\!\perp X_c | X_a$ se, e somente se,

1. $cov(X_c, X_b | X_a) = 0$, ou $V_{bc/a} = V_{bc} - V_{ba}V_{aa}^{-1}V_{ac} = 0$; ou
2. o bloco da inversa variância $D_{bc} = 0$.

Em particular, se X_b e X_c são ambos unidimensionais, então:

$$X_i \perp\!\!\!\perp X_j | X_{k \setminus \{i,j\}} \text{ se, e somente se, } d_{ij} = 0.$$

Modelos Grafos Gaussianos

Dado um grafo independente G e um vetor aleatório $X = (X_1, X_2, \dots, X_k)$ k - dimensional, um modelo grafo Gaussiano é uma família de distribuições normais para X , que satisfaz a independência condicional para pares de variáveis no grafo independente. O Corolário 2.3 mostra que tal independência condicional é equivalente a especificar zeros nos parâmetros da inversa da variância e corresponde à ausência de uma aresta em G .

Exemplo 2.4. : Considere o vetor aleatório X de dimensão quatro, $k = 4$, e um modelo grafo Gaussiano para X é tal que a inversa da variância $D = var(X_1, X_2, X_3, X_4)^{-1}$ é da forma

$$D = \begin{pmatrix} d_{11} & d_{12} & 0 & 0 \\ d_{21} & d_{22} & d_{23} & 0 \\ 0 & d_{23} & d_{33} & d_{34} \\ 0 & 0 & d_{34} & d_{44} \end{pmatrix}$$

Desta forma, o grafo de X é dado pela Figura A.2

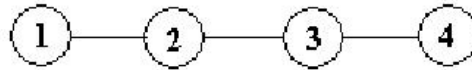


FIGURA 2.18: Grafo não-orientado com 4 vértices.

A propriedade de Markov global afirma que $X_1 \perp\!\!\!\perp X_3 | X_2$ e conseqüentemente, concluímos que $var(X_1, X_2, X_3)^{-1}$ é da forma

$$\begin{pmatrix} * & * & 0 \\ * & * & * \\ 0 & * & * \end{pmatrix}$$

independente dos valores de não-zeros (*) em D .

2.3.2 Função de Verossimilhança

Uma amostra de N vetores aleatórios independentes e identicamente distribuídos x^1, x^2, \dots, x^N , cada um com função densidade Normal multivariada k -dimensional $f(x; \theta)$, com parâmetro θ , tem função log-verossimilhança

$$l(\theta) = l(\theta; x^1, x^2, \dots, x^N) = \sum_{l=1}^N \log f(x^l; \theta).$$

Como o objetivo é estimar a estrutura de dependência e independência condicional existente entre os vetores, temos interesse nos elementos da matriz de variâncias V , denotada por $V = \{v_{ij}\}$, e os elementos da variância inversa $D = V^{-1} = \{d_{ij}\}$.

A fórmula da função log-verossimilhança para a função densidade normal multivariada é

$$\begin{aligned} l(\mu, V) &= \sum_{l=1}^N \log \left((2\pi)^{-\frac{k}{2}} \det(D)^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(x^l - \mu)^T D (x^l - \mu)] \right\} \right) \\ &= \log(2\pi)^{-\frac{Nk}{2}} + N \log \left(\det(D)^{\frac{1}{2}} \right) - \frac{1}{2} \sum_{l=1}^N [(x^l - \mu)^T D (x^l - \mu)] \end{aligned}$$

como $D = V^{-1}$, temos

$$= -\frac{1}{2} \log(2\pi)^{Nk} - \frac{1}{2} N \log(\det V) - \frac{1}{2} \sum_{l=1}^N [(x^l - \mu)^T V^{-1} (x^l - \mu)]$$

$\Rightarrow 2l(\mu, V) = -\log(2\pi)^{Nk} - \sum_{l=1}^N [(x^l - \mu)^T V^{-1} (x^l - \mu)] - N \log(\det V)$; sendo $(\log(2\pi)^{Nk}) = \text{const}$, temos

$$2l(\mu, V) = \text{const} - \sum_{l=1}^N [(x^l - \mu)^T V^{-1} (x^l - \mu)] - N \log(\det V)$$

A constante na função log-verossimilhança não depende das observações nem dos parâmetros em V . Definimos o vetor de médias amostrais por $\bar{x} = N^{-1}(\sum x^l)$ e a variância amostral por $S = \{s_{ij}\}$, onde o elemento

$$s_{ij} = \frac{1}{N} \sum_{l=1}^N (x_i^l - \bar{x}_i)(x_j^l - \bar{x}_j)^T$$

é a covariância amostral da i -ésima e j -ésima coordenadas de X .

Primeiramente, notemos que

$$\sum [(x^l - \mu)^T V^{-1} (x^l - \mu)] = \sum [(x^l - \bar{x})^T (x^l - \bar{x})] + N [(\bar{x} - \mu)^T V^{-1} (\bar{x} - \mu)],$$

porque o termo do produto cruzado $\sum [(x^l - \bar{x})^T (\bar{x} - \mu)]$, é zero.

A primeira parte da soma de quadrados pode ser substancialmente simplificada usando

$$v^T V^{-1} v = \text{tr}(v^T V^{-1} v) = \text{tr}(V^{-1} v v^T),$$

o qual funciona porque o produto interno é um escalar e $\text{tr}(AB) = \text{tr}(BA)$, para matrizes A e B genéricas.

Portanto,

$$\begin{aligned} \sum [(x^l - \bar{x})^T V^{-1} (x^l - \bar{x})] &= \sum \text{tr}(V^{-1} (x^l - \bar{x})(x^l - \bar{x})^T) \\ &= \text{tr} \left(V^{-1} \sum (x^l - \bar{x})(x^l - \bar{x})^T \right) \end{aligned}$$

assim como $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B) = N \text{tr}(V^{-1} S)$.

Assim, temos:

$$2l(\mu, V) = \text{const} - N [(\bar{x} - \mu)^T V^{-1} (\bar{x} - \mu)] - N \text{tr}(V^{-1} S) - N \log(\det V).$$

Proposição 2.2. (*A função verossimilhança e a divergência*). A função log-verossimilhança para a média μ e variância V , da distribuição normal multivariada baseada na amostra aleatória de N observações, podem ser escritas como:

1. $2l(\mu, V) = \text{const} - N [(\bar{x} - \mu)^T V^{-1} (\bar{x} - \mu)] - N \text{tr}(V^{-1} S) - N \log(\det V)$; ou
2. $l(\mu, V) = \text{const} - NI(f^{\bar{x}, S}; f^{\mu, V})$;

onde $I(f^{\bar{x}, S}; f^{\mu, V})^4$ é a informação divergência ² entre duas densidades normais multivariadas (ver Whittaker, 1990, página 172).

Corolário 2.4. (*Maximização não restrita*). Quando V é sem restrição, os estimadores de máxima verossimilhança únicos da média μ e variância V , para uma amostra de observações normais independentes são a média amostral \bar{x} e variância amostral S .

Demonstração. Como a divergência é não-negativa, o valor mínimo de $I(f^{\bar{x}, S}; f^{\mu, V})$ é 0. Isto ocorre quando $\mu = \bar{x}$ e $V = S$. Isto é único porque a divergência é definida positiva tal que se $I(f; g) = 0$ então $f = g$, e conseqüentemente $f^{\bar{x}, S}(x) = f^{\mu, V}(x)$ para todo x . Disto segue que $\mu = \bar{x}$ e $V = S$. □

Podemos escrever a função de verossimilhança para um modelo grafo dada uma amostra aleatória da distribuição normal multivariada, a próxima tarefa é verificar este máximo.

O caso geral

Consideramos agora o caso geral e o próximo teorema relata isso. A prova pode ser vista em Speed e Kiiveri (1986) usando o trabalho de Csiszar (1975).

Teorema 2.3. (*Equações de verossimilhança para modelos grafos Gaussianos*). O estimador de máxima verossimilhança de um modelo grafo G , baseado em uma amostra aleatória da distribuição Normal multivariada, satisfaz as equações de verossimilhança

$$\hat{d}_{ij} = 0,$$

quando vértices i e j são não adjacentes em G , e,

$$\hat{V}_{aa} = S_{aa}$$

quando o subconjunto a de vértices em G formam um clique. Os parâmetros D e V são relacionados por $\hat{D} = \hat{V}^{-1}$, e são únicos com probabilidade um.

²Segue no Apêndice A a proposição sobre informação divergência

2.4 Um Exemplo

A Tabela 2.1 mostra notas de 88 estudantes em cinco disciplinas: mecânica, vetores, álgebra, análise e estatística. Todas as variáveis são medidas na mesma escala, e não há valores desconhecidos ou omitidos. Estes dados são retirados de Mardia, Kent e Bibby (1979).

TABELA 2.1: Notas em cinco matérias para 88 estudantes

me	ve	al	an	es	me	ve	al	an	es	me	ve	al	an	es
77	82	67	67	81	30	69	50	52	45	62	44	36	22	42
63	78	80	70	81	46	49	53	59	37	48	38	41	44	33
75	73	71	66	81	40	27	54	61	61	34	42	50	47	29
55	72	63	70	68	31	42	48	57	68	18	51	40	56	30
63	63	65	70	63	36	59	51	45	51	35	36	46	48	29
53	61	72	64	73	56	40	56	54	35	59	53	37	22	19
51	67	65	65	68	46	56	57	49	32	41	41	43	30	33
59	70	68	62	56	45	42	55	56	40	31	52	37	56	30
62	60	58	62	70	42	60	54	49	33	17	51	52	35	31
64	72	60	62	45	40	63	53	54	25	34	30	50	47	36
52	64	60	63	54	23	55	59	53	44	46	40	47	29	17
55	67	59	62	44	48	48	49	51	37	10	46	36	47	39
50	50	64	55	63	41	63	49	46	34	46	37	45	15	30
65	63	58	56	37	46	52	53	41	40	30	34	43	46	18
31	55	60	57	73	46	61	46	38	41	13	51	50	25	31
60	64	56	54	40	40	57	51	52	31	49	50	38	23	9
44	69	53	53	53	49	49	45	48	39	18	32	31	45	40
42	69	61	55	45	22	58	53	56	41	8	42	48	26	40
62	46	61	57	45	35	60	47	54	33	23	38	36	48	15
31	49	62	63	62	48	56	49	42	32	30	24	43	33	25
44	61	52	62	46	31	57	50	54	34	3	9	51	47	40
49	41	61	49	64	17	53	57	43	51	7	51	43	17	22
2	58	61	63	67	49	57	47	39	26	15	40	43	23	18
49	53	49	62	47	59	50	47	15	46	15	38	39	28	17
54	49	56	47	53	37	56	49	28	45	5	30	44	36	18
54	53	46	59	44	40	43	48	21	61	12	30	32	35	21
44	56	55	61	36	35	35	41	51	50	5	26	15	20	20
18	44	50	57	81	38	44	54	47	24	0	40	21	9	14
46	52	65	50	35	43	43	38	34	49					
32	45	49	57	64	39	46	46	32	43					

Construindo os gráficos de ramo-e-folhas para cada matéria, temos que as distribuições marginais das notas são bem aproximadas pela distribuição normal.

Mecânica

0	035578
1	02235577888
2	233
3	000111112445556789
4	00001122344456666666888999999
5	012344556999
6	02223345
7	57

Vetores

0	9
1	
2	467
3	00024567888
4	0000112222334445666899999
5	00011112223333556666777889
6	00011133334477999
7	02238
8	2

Álgebra

1	5
2	1
3	1266677889
4	01133333455666677778889999999
5	000000111223333344455666778899
6	000111123455578
7	12
8	0

Análise

0	9
1	557
2	012233567889
3	023556789
4	1234556677777888999
5	0112233444444556666777799
6	112222223334567
7	000

Estatística

0	9
1	45778889
2	012455699
3	0011122333344556677799
4	00000001123444555556679
5	0113346
6	11233447888
7	033
8	1111

FIGURA 2.19: Gráficos ramo-e-folhas de cinco disciplinas.

Além dos gráficos de ramo-e-folhas, os diagramas de dispersão nos dão algumas informações usuais do relacionamento entre as notas das matérias: eles indicam as correlações entre essas notas e sugerem normalidade para a distribuição bivariada de notas.

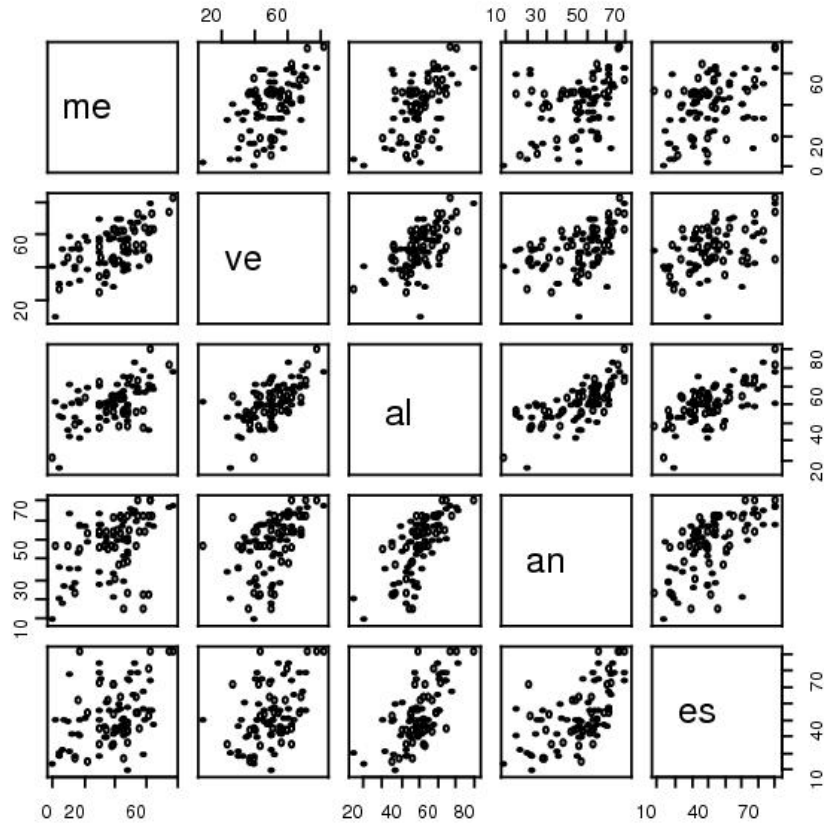


FIGURA 2.20: Diagramas de dispersão de pares de disciplinas.

A matriz correlação

A matriz de correlações é obtida da matriz de variâncias pelo escalonamento de linhas e colunas de tal forma que as entradas diagonais sejam todas unitárias. Segue a matriz de correlações

$$\begin{array}{c}
 \begin{array}{ccccc}
 & me & ve & al & an & es \\
 me & \left(\begin{array}{ccccc}
 1.0000000 & 0.5534052 & 0.5467511 & 0.4106496 & 0.3890993 \\
 0.5534052 & 1.0000000 & 0.6096447 & 0.4845900 & 0.4364487 \\
 0.5467511 & 0.6096447 & 1.0000000 & 0.7092947 & 0.6647357 \\
 0.4106496 & 0.4845900 & 0.7092947 & 1.0000000 & 0.6091436 \\
 0.3890993 & 0.4364487 & 0.6647357 & 0.6091436 & 1.0000000
 \end{array} \right)
 \end{array}
 \end{array}$$

O que não é bem conhecido é como interpretar os elementos da matriz inversa da variância e matriz de correlações. Primeiramente e mais importante é se um elemento da matriz inversa da variância for zero então as variáveis correspondentes são condicionalmente independentes dado as variáveis restantes. Isso imediatamente motiva o cálculo da

inversa da variância. A experiência mostra que é mais fácil interpretar uma correlação do que uma covariância. Portanto, a matriz de correlações inversas é calculada

$$\begin{array}{c}
 \\
 \\
 \\
 \\
 \\
 \end{array}
 \begin{array}{ccccc}
 & me & ve & al & an & es \\
 me & \left(\begin{array}{ccccc}
 1.603629934 & -0.55933763 & -0.5056822 & -0.003467434 & -0.04159189 \\
 -0.559337635 & 1.80199366 & -0.6593609 & -0.153174183 & -0.03723407 \\
 -0.505682205 & -0.65936092 & 3.0282494 & -1.096598544 & -0.86046180 \\
 -0.003467434 & -0.15317418 & -1.0965985 & 2.174888501 & -0.52766934 \\
 -0.041591891 & -0.03723407 & -0.8604618 & -0.527669341 & 1.92584025
 \end{array} \right)
 \end{array}$$

Cada elemento diagonal da matriz de correlações inversa está relacionado com a proporção de variação da variável correspondente explicado pela regressão das variáveis restantes. Mas explicitamente cada elemento da diagonal é igual a $1/(1 - R^2)$ onde R é o coeficiente de correlação múltiplo entre essa variável e o resto. Então, escalonando a inversa da matriz de correlações para se obter a unidade na diagonal, temos

$$\begin{array}{c}
 \\
 \\
 \\
 \\
 \\
 \end{array}
 \begin{array}{ccccc}
 & me & ve & al & an & es \\
 me & \left(\begin{array}{ccccc}
 1.000000000 & -0.32903745 & -0.2294720 & -.001856682 & -0.02366718 \\
 -0.329037453 & 1.00000000 & -0.2822609 & -0.077373180 & -0.01998728 \\
 -0.229472038 & -0.28226092 & 1.0000000 & -0.427300445 & -0.35630806 \\
 -0.001856682 & -0.07737318 & -0.4273004 & 1.000000000 & -0.25782987 \\
 -0.023667177 & -0.01998728 & -0.3563081 & -0.257829872 & 1.00000000
 \end{array} \right)
 \end{array}$$

Os elementos que não estão na diagonal da matriz de correlações inversa escalonada são os negativos dos coeficientes de correlações parciais entre os pares correspondentes de variáveis dadas as variáveis restantes.

Ao contrário da matriz correlação original das notas, há uma estrutura identificável nesta matriz de correlação inversa escalonada: os elementos do lado esquerdo abaixo, são todos próximos a zero, o que sugere que a matriz pode ser aproximada pela estrutura

$$\begin{array}{c}
 \\
 \\
 \\
 \\
 \\
 \end{array}
 \begin{array}{ccccc}
 & me & ve & al & an & es \\
 me & \left(\begin{array}{ccccc}
 * & & & & \\
 * & * & & & \\
 * & * & * & & \\
 0 & 0 & * & * & \\
 0 & 0 & * & * & *
 \end{array} \right)
 \end{array}$$

onde * denota valores diferentes de zero.

Desta forma, o grafo independente das notas, onde os vértices representam as matérias, é mostrado na Figura 2.21

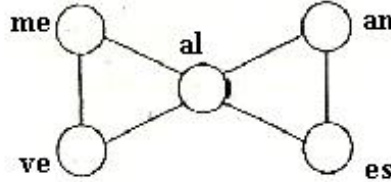


FIGURA 2.21: Modelo grafo para as notas.

Percebemos, por exemplo, que a correlação parcial entre mecânica e análise é zero então estas variáveis são condicionalmente independentes dadas as variáveis, vetores, álgebra e estatística. Mas, visto que álgebra separa mecânica e análise conclui-se que estas duas variáveis são condicionalmente independentes dado álgebra somente.

Verificamos que as propriedades de Markov estão inclusas no grafo:

- *Propriedade de Markov para pares de variáveis:*

$$me \perp\!\!\!\perp an | (al, ve, es);$$

$$ve \perp\!\!\!\perp es | (al, me, an);$$

$$me \perp\!\!\!\perp es | (al, an, ve);$$

$$ve \perp\!\!\!\perp an | (al, me, es).$$

- *Propriedade de Markov local:*

$$me \perp\!\!\!\perp (an, es) | (ve, al);$$

$$ve \perp\!\!\!\perp (an, es) | (me, al);$$

$$an \perp\!\!\!\perp (me, ve) | (al, es);$$

$$es \perp\!\!\!\perp (me, ve) | (al, an).$$

- *Propriedade de Markov Global:* Consideremos os subconjuntos $b = \{me, ve\}$, $c = \{an, es\}$ e $a = \{al\}$, desta forma $X_b \perp\!\!\!\perp X_c | X_a$, como vemos no grafo.

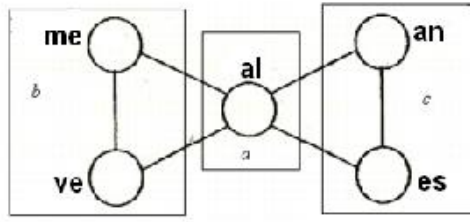


FIGURA 2.22: Modelo grafo para a Propriedade de Markov global.

Com isso, temos um exemplo da equivalência das propriedades de Markov.

Estimadores de máxima verossimilhança

De Whittaker (1990), o estimador de máxima verossimilhança para a variância é $\hat{V}_{aa} = S_{aa}$, onde a denota o subconjunto de vértices em G que formam um clique, a matriz variância amostral,

$$S = \begin{pmatrix} 305.77 & & & & & \\ 127.22 & 172.84 & & & & \\ 101.58 & 85.16 & 112.88 & & & \\ 106.41 & 94.41 & 111.68 & 219.61 & & \\ 117.40 & 99.01 & 121.87 & 155.77 & 297.75 & \end{pmatrix}$$

Os cliques no grafo são: $\{me, al\}$, $\{me, ve\}$, $\{ve, al\}$, $\{al, an\}$, $\{an, es\}$ e $\{al, es\}$ e suas equações de verossimilhança implicam que:

$$\begin{pmatrix} \hat{v}_{meme} \\ \hat{v}_{meal} & \hat{v}_{alal} \end{pmatrix} = \begin{pmatrix} 305.77 \\ 101.58 & 112.88 \end{pmatrix}; \quad \begin{pmatrix} \hat{v}_{meme} \\ \hat{v}_{meve} & \hat{v}_{veve} \end{pmatrix} = \begin{pmatrix} 305.77 \\ 127.22 & 172.84 \end{pmatrix};$$

$$\begin{pmatrix} \hat{v}_{veve} \\ \hat{v}_{veal} & \hat{v}_{alal} \end{pmatrix} = \begin{pmatrix} 172.84 \\ 85.16 & 112.88 \end{pmatrix}; \quad \begin{pmatrix} \hat{v}_{alal} \\ \hat{v}_{alal} & \hat{v}_{anan} \end{pmatrix} = \begin{pmatrix} 112.88 \\ 111.68 & 219.61 \end{pmatrix};$$

$$\begin{pmatrix} \hat{v}_{anan} \\ \hat{v}_{anes} & \hat{v}_{eses} \end{pmatrix} = \begin{pmatrix} 219.61 \\ 155.77 & 297.75 \end{pmatrix}; \quad \begin{pmatrix} \hat{v}_{alal} \\ \hat{v}_{ales} & \hat{v}_{eses} \end{pmatrix} = \begin{pmatrix} 112.88 \\ 121.87 & 297.75 \end{pmatrix};$$

que podem ser juntadas, de tal forma que fiquem assim,

$$\begin{pmatrix} \hat{v}_{meme} \\ \hat{v}_{meve} & \hat{v}_{veve} \\ \hat{v}_{meal} & \hat{v}_{veal} & \hat{v}_{alal} \\ \hat{v}_{mean} & \hat{v}_{vean} & \hat{v}_{alan} & \hat{v}_{anan} \\ \hat{v}_{mees} & \hat{v}_{vees} & \hat{v}_{ales} & \hat{v}_{anes} & \hat{v}_{eses} \end{pmatrix} = \begin{pmatrix} 305.77 \\ 127.22 & 172.84 \\ 101.58 & 85.16 & 112.88 \\ ? & ? & 111.68 & 219.61 \\ ? & ? & 121.87 & 155.77 & 297.75 \end{pmatrix}.$$

Agrupando com as equações correspondentes para a independência no grafo, e acrescentando os termos $\hat{d}_{mean} = 0$, $\hat{d}_{vean} = 0$, $\hat{d}_{mees} = 0$ e $\hat{d}_{vees} = 0$, obtemos os estimadores de máxima verossimilhança.

Podemos perceber neste modelo grafo que álgebra e análise são suficientes para prever a nota de estatística, que álgebra e vetores são suficientes para prever mecânica, mas que são necessárias as outras quatro notas para prever álgebra. Por esta razão modelos grafos são essenciais na visualização das dependências e independências condicionais, quando trabalhamos com um número grande de variáveis e as propriedades de Markov, auxiliam na identificação desses condicionamentos.

2.5 Modelos Grafos: Caso Discreto

Nesta seção descrevemos modelos grafos apropriados para tabelas de contingência multi-entradas, baseadas na distribuição Multinomial. Edwards (1995) mostra que existem subclasses de modelos loglineares hierárquicos, especificados pela parametrização da função densidade em termos dos coeficientes da expansão loglinear, os termos u . O desenvolvimento desta parametrização para tabelas de probabilidades, conhecidas como modelos loglineares, foi iniciado por Birch (1963), Godman (1970), Haberman (1974), e Bishop, Fienberg e Holland (1975).

A modelagem loglinear tem demonstrado ser uma técnica confiável de extração de informação de tabelas de contingência complexas. Além disso, apresentamos o Paradoxo de Simpson, que estuda a troca de sentido de associação entre variáveis (numéricas ou qualitativas) quando se controla o efeito de uma terceira variável. E por fim, concluímos com um exemplo, onde é possível identificar o paradoxo de Simpson.

2.5.1 Tabelas de Três Entradas

Os modelos log-lineares são assim chamados porque eles usam modelos lineares para logaritmos da contagem de elementos esperados. Consideramos N observações de três variáveis discretas A , B e C . Tomando A com $\#A$ níveis, B com $\#B$ níveis e C com $\#C$ níveis, formamos uma tabela de três entradas de contagem por classificação-cruzada A , B e C e denotamos uma contagem típica com n_{jkl} onde j pode tomar os valores $1, \dots, \#A$, k pode ser $1, \dots, \#B$ e l pode ser da forma $1, \dots, \#C$. Similarmente, escrevemos a probabilidade da célula (j, k, l) da tabela - como p_{jkl} e o valor esperado como $m_{jkl} = Np_{jkl}$, com N sendo o número de observações.

Este planejamento amostral, no qual o número de observações é fixado, é chamado amostragem multinomial. Outros planejamentos podem ser usados: por exemplo, na amostragem linha-multinomial, o número de observações em cada etapa, digamos A , é fixado com antecedência. Um modelo simplificado para uma tabela de três entradas escreve o logaritmo da probabilidade de cada célula como sendo

$$\ln(p_{jkl}) = u + u_j^A + u_k^B + u_l^C \quad (2.1)$$

onde os u 's são os parâmetros, também chamados de *termos de interação*.

O modelo pode ser re-formulado como $p_{jkl} = p_{j++}p_{+k+}p_{++l}$ onde os $+$'s denotam ocultamento dos respectivos índices.

Em outras palavras, no modelo (2.1) A , B e C são independentes.

Para identificar o modelo, uma fórmula pode ser usada. Este consiste de uma lista de termos chamados de geradores, que correspondem à interação maximal no modelo. Para o modelo independente, a fórmula é A, B, C e o grafo independente é mostrado na Figura 2.23

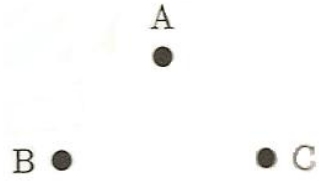


FIGURA 2.23: Grafo para variáveis A, B, C independentes.

Um modelo log-linear mais complexo pode ser escrito

$$\ln(p_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jk}^{AB} + u_{jl}^{AC}.$$

Desde que u_{jk}^{AB} e u_{jl}^{AC} sejam *termos de interação* maximal, o modelo será AB, AC.

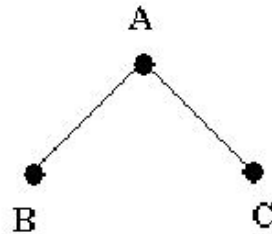
O modelo pode ser reformulado em termos das probabilidades de elementos como:

$$p_{jkl} = \frac{p_{jk+}p_{j+l}}{p_{j++}} \quad \text{ou} \quad \frac{p_{jkl}}{p_{j++}} = \left(\frac{p_{jk+}}{p_{j++}} \right) \left(\frac{p_{j+l}}{p_{j++}} \right)$$

o que equivale a

$$P_r(B = k, C = l | A = j) = P_r(B = k | A = j)P_r(C = l | A = j).$$

Esta última forma revela que B e C são condicionalmente independentes, e denotamos por $B \perp\!\!\!\perp C | A$, dado que $A=j$. O grafo do modelo é mostrado na Figura 2.24

FIGURA 2.24: Grafo de $B \perp\!\!\!\perp C | A$.

O modelo log-linear geral para três fatores pode ser escrito como:

$$\ln(p_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jk}^{AB} + u_{jl}^{AC} + u_{kl}^{BC} + u_{jkl}^{ABC}$$

correspondendo ao modelo fórmula ABC. Esta não impõe restrições no p_{jkl} . O modelo prévio AB, AC pode ser visto como um caso especial do modelo geral definido pela posição $u^{BC} = u^{ABC} = 0$, isto é, a posição $u_{kl}^{BC} = 0$ para todo k e l e $u_{jkl}^{ABC} = 0$ para todo j, k e l .

Similarmente, o modelo independente é o caso especial do modelo geral com $u^{BC} = u^{AC} = u^{AB} = u^{ABC} = 0$.

Quase invariavelmente, a atenção é restrita para modelos loglineares *hierárquicos*. Aqui, o termo hierárquico indica que se um termo é zero, então todos de ordem superior relativos são também zero. Por exemplo, se no nosso conjunto $u^{AB} = 0$, mas admitimos não zeros no termo de interação de três fatores u_{jkl}^{ABC} , então este poderia definir um modelo não-hierárquico. Modelos não-hierárquicos são difíceis para interpretar, e por esta razão, eles não têm sido muito usados na prática.

Função de Verossimilhança

Sob amostragem de N observações da distribuição multinomial, a verossimilhança de uma dada tabela é

$$L(\{p_{jkl}\} | \{n_{jkl}\}) = \frac{N!}{\prod_{jkl} n_{jkl}!} \prod_{jkl} p_{jkl}^{n_{jkl}}.$$

Os valores de p_{jkl} que maximizam esta expressão para um dado modelo são os estimadores de máxima verossimilhança (EMV), escritos \hat{p}_{jkl} . Como a função logaritma é monótona, os EMV também maximizam a log-verossimilhança,

$$l(\{p_{jkl}\} | \{n_{jkl}\}) = \ln \left(\frac{N!}{\prod_{jkl} n_{jkl}!} \right) + \sum_{jkl} n_{jkl} \ln(p_{jkl}).$$

Deviance

A *deviance* de um modelo M_0 é uma função da estatística de teste de razão de verossimilhança de M_0 versus o modelo M_f saturado ou sem restrição, tal que

$$Dev = 2(\widehat{l}_f - \widehat{l}_0)$$

onde \widehat{l}_f e \widehat{l}_0 são log-verossimilhanças maximizadas sob M_f e M_0 , respectivamente. A *deviance* é uma medida conveniente de M_0 apropriado, devido as três propriedades:

1. Quando $M_0 = M_f$, a Dev é zero;
2. Se M_0 é verdadeira, $Dev \sim \chi^2_{(k)}$ assintoticamente onde os k graus de liberdade (k) são dados pela diferença do número de parâmetros livres entre M_0 e M_f . Sob M_0 , $E(Dev) = k$;
3. Quando M_0 não é verdadeira, a distribuição assintótica da *deviance* é estatisticamente maior que $\chi^2_{(k)}$, isto é, $E(Dev) > k$.

Sob M_f , os EMV de p_{jkl} são $\frac{n_{jkl}}{N}$ tal que a *deviance* pode ser escrita como

$$d^2 = 2 \sum_{jkl} n_{jkl} \ln(n_{jkl}/N) - \sum_{jkl} n_{jkl} \ln(\widehat{p}_{jkl}) = 2 \sum_{jkl} n_{jkl} \ln \left(\frac{n_{jkl}}{\widehat{m}_{jkl}} \right)$$

onde os \widehat{p}_{jkl} são EMV de p_{jkl} sob M_0 , e $\widehat{m}_{jkl} = N\widehat{p}_{jkl}$. Esta expressão para a *deviance* é também válida sob outros planejamentos amostrais amostrais, por exemplo, independência - Poisson ou amostragem linha multinomial.

Para dois modelos encaixados, $M_0 \subseteq M_1$, a *deviance* é dada como

$$d = 2 \sum_{jkl} n_{jkl} \ln \left(\frac{\widehat{m}_{jkl}^1}{\widehat{m}_{jkl}^0} \right)$$

onde \widehat{m}_{jkl}^0 e \widehat{m}_{jkl}^1 são EMV sob M_0 e M_1 , respectivamente. Sob M_0 , assintoticamente d é $\chi^2_{(k)}$, onde k é a diferença no número de parâmetros livres entre M_0 e M_1 .

Sob o modelo sem restrição, temos

$$p_{jkl} = \exp \left(u + u_j^A + u_k^B + u_l^C + u_{jk}^{AB} + u_{jl}^{AC} + u_{kl}^{BC} + u_{jkl}^{ABC} \right)$$

Para o modelo AB, AC, que é definido pela posição $u^{BC} = 0$, e daí $u^{ABC} = 0$, a probabilidade pode ser fatorada em

$$p_{jkl} = \exp(u + u_j^A + u_k^B + u_{jk}^{AB}) \exp(u_l^C + u_{jl}^{AC}),$$

tal que o primeiro termo não envolve C e o segundo termo não envolve B. Então, do critério da fatoração segue que $B \perp\!\!\!\perp C | A$. Geralmente, sob qualquer modelo hierárquico, dois fatores são condicionalmente independentes dado o resto se, e somente se, o correspondente termo de interação de dois fatores é zero.

Este resultado forma a base de um subconjunto de modelos hierárquicos. As interações de ordem superior incluídas no modelo são especificadas pelo termo de interação de dois fatores. A característica interessante de tais modelos é que eles podem ser interpretados somente em termos da independência condicional.

Por exemplo, o modelo AB, AC é grafo desde que u^{BC} seja zero. O exemplo simples de um modelo grafo não-hierárquico é AB, BC, AC, o qual o conjunto de interação u^{ABC} de três fatores é zero. Temos então, a Figura 2.25

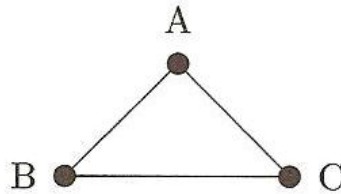


FIGURA 2.25: Grafo não hierárquico AB, BC, AC.

O modelo grafo correspondente a este grafo é o modelo saturado ABC.

Exemplo 2.5. (*Comportamento de ninho de lagartos*): Consideremos a tabela de três-entradas de Fienberg (1980). A Tabela 2.2 consiste de dados de comportamentos nos ninhos de duas espécies de lagartos. Para uma amostra de 164 lagartos, as seguintes variáveis são dadas: espécie (A), diâmetro dos ninhos (B): 1 = estreito, 2 = largo, e altura do ninho (C): 1 = alto, 2 = baixo. Os dados originais, diâmetro e altura do ninho são contínuos, mas dicotomizamos. Os dados estão mostrados na Tabela 2.2

TABELA 2.2: Classificação cruzada por sexo e admissão

Espécies	Diâmetro Ninho	Peso Ninho (> 4.75)	Peso Ninho (< 4.75)
Anoli	≤ 4	32	86
Anoli	≥ 4	11	35
Distichus 2	≤ 4	61	73
Distichus 2	≥ 4	41	70

De acordo com Edwards (1995), é necessário excluir a aresta BC do grafo completo, resultando no modelo AC, AB, e com a ajuda da *deviance*, vemos que este é um modelo aceitável. O grafo resultante é dado pela Figura 2.26

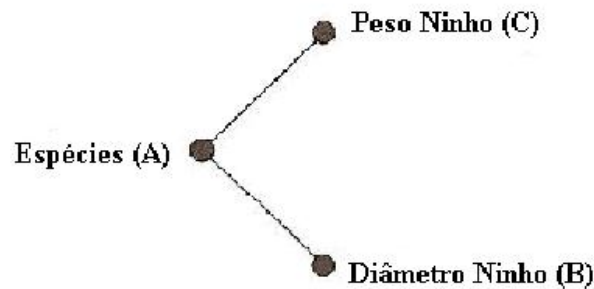


FIGURA 2.26: Grafo resultante no modelo AC, AB.

Este grafo nos diz que diâmetro e peso dos ninhos são condicionalmente independentes dada espécie, ou seja, $B \perp\!\!\!\perp C | A$.

2.5.2 Tabelas Multi-Entradas

Nesta seção descrevemos os modelos log-lineares para tabelas multi-entradas. Entender os resultados da seção anterior para tabelas de múltiplas entradas é essencialmente simples; a principal dificuldade é notacional. Uma vez que para tabelas de três entradas o fator nome e níveis podem ser escritos explicitamente, assim por exemplo em p_{jkl}^{ABC} , u_{kl}^{BC} e n_{j+l} , este método torna-se muito pesado quando o número de fatores é arbitrário, e por isso, precisamos de uma notação mais geral.

Seja Δ um conjunto de variáveis discretas. Supomos que os dados para serem analisados consistem de N observações e p variáveis, as quais podem ser resumidas em uma tabela p -dimensional de contagem, formada pela classificação cruzada das p variáveis discretas.

Escrevemos uma célula típica na tabela como p -upla

$$i = (i_1, \dots, i_p)$$

e denotamos o número de observações na célula i como n_i , a probabilidade da célula i como p_i , e a esperança de célula contagem como $m_i = Np_i$. Seja I o conjunto de todas as células da tabela. Consideramos somente tabelas completas, onde o número de células em I é o número de níveis dos fatores em Δ .

Sob amostragem multinomial, a verossimilhança de uma dada tabela $\{n_i\}_{i \in I}$ é

$$L(\cdot) = \frac{N!}{\prod_{i \in I} n_i!} \prod_{i \in I} p_i^{n_i}.$$

Note que precisamos de uma notação para células marginais. Para uma célula $i \in I$ e um subconjunto $a \subseteq \Delta$, seja i_a o correspondente a sub p -upla de i , e seja I_a o conjunto de todos os possíveis i_a .

Escrevemos um termo de interação geral como u_i^a , onde é entendido que u_i^a depende de i somente através de i_a . Então, o modelo loglinear saturado para p fatores pode ser escrito como

$$\ln(p_i) = \sum_{a \subseteq \Delta} u_i^a.$$

Para este modelo, os EMV's são $\hat{p}_i = \frac{n_i}{N}$.

Um modelo pode ser especificado através de

$$d_1, \dots, d_r,$$

onde os conjuntos $d_i \subseteq \Delta$ são chamados *geradores*. Estes identificam as interações maximais que não são zero. Então, nós escrevemos o modelo como

$$\ln(p_i) = \sum_{a \subseteq \Delta: a \subseteq d_j \text{ para algum } j} u_i^a. \quad (2.2)$$

Equações de Verossimilhança

Seja $\{n_{i_a}\}_{i_a \in I_a}$ a tabela marginal de contagem observada correspondente a a e seja $\{\hat{m}_{i_a}\}_{i_a \in I_a}$ a tabela marginal de contagem ajustada para alguma estimativa $\{\hat{m}_i\}_{i \in I}$ de $\{m_i\}_{i \in I}$.

Considere um dado modelo com fórmula d_1, \dots, d_r . Um conjunto minimal de estatísticas suficientes é dado pelo conjunto de tabelas marginais $\{n_{i_a}\}_{i_a \in I_a}$ correspondentes aos geradores no modelo, isto é, para $a = d_1, \dots, d_r$. Pela equiparação, a estatística suficiente minimal destes valores esperados sob o modelo, formam as equações de verossimilhança

$$\{n_{i_a}\}_{i_a \in I_a} = \{\hat{m}_{i_a}\}_{i_a \in I_a}, \quad (2.3)$$

para $a = d_1, \dots, d_r$. Isto é, os EMV's \hat{m}_i sob o modelo são dados como solução para (2.3) que também satisfazem o modelo de restrições como expressado em (2.2).

Deviance

A deviance sob M_0 é

$$d^2 = 2 \sum_{i \in I} n_i \ln \left(\frac{n_i}{\hat{m}_i^0} \right),$$

e a deviance diferença para $M_0 \subseteq M_1$ é

$$\begin{aligned} d &= d_0^2 - d_1^2 = 2 \sum_{i \in I} n_i \ln \left(\frac{n_i}{\hat{m}_i^0} \right) - 2 \sum_{i \in I} n_i \ln \left(\frac{n_i}{\hat{m}_i^1} \right) \\ &= 2 \sum_{i \in I} n_i \ln \left(\frac{\hat{m}_i^1}{\hat{m}_i^0} \right), \end{aligned}$$

onde $\{\hat{m}_i^0\}_{i \in I}$ são células de contagem ajustadas sob M_0 e M_1 , respectivamente. Sob M_0 , d tem uma distribuição assintótica χ^2 com graus de liberdade dados pela diferença em número de parâmetros livres entre M_0 e M_1 .

2.5.3 Grafos e Modelos

Dado um modelo, formamos o grafo pela conexão de todos os pares de vértices que aparecem no mesmo gerador. Por exemplo, o grafo ABC, BCD, DE, BE é

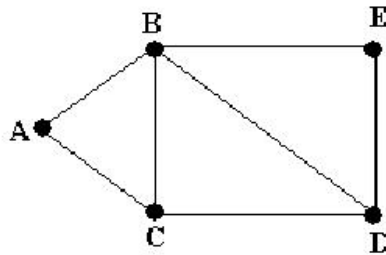


FIGURA 2.27: Grafo correspondente ao modelo ABC, BCD, DE, BE.

Inversamente, dado um grafo, podemos construir o modelo grafo correspondente pela descoberta de cliques do grafo. Por exemplo, os cliques do grafo mostrado na Figura 2.27 são $\{A, B, C\}$, $\{B, C, D\}$ e $\{B, D, E\}$, assim, o modelo correspondente tem fórmula ABC, BCD, BDE.

Modelos para os quais as fórmulas explícitas para os EMV's podem ser derivados são chamados decomponíveis. Darroch, Lauritzen, e Speed (1980) caracterizam modelos loglineares decomponíveis como modelos grafos para os quais os grafos são triangulares, isto é, não contém ciclos de grande comprimento.

2.6 Paradoxo de Simpson

Está provado pelo paradoxo de Simpson (Simpson, 1951) que para analisar dependências entre k variáveis não é suficiente estudar as dependências entre pares de variáveis. Para ilustrar esse fato, baseados em Edwards (1995), supomos possível discriminação sexual em relação as admissões em uma universidade. A Tabela 2.3 mostra alguns dados fictícios para uma dada universidade. Parece haver diferença contra candidatas

mulheres, pois somente 41,6% dessas são admitidas contra 58,3% de homens.

TABELA 2.3: Classificação cruzada por sexo e admissão

Sexo	Nº Total	Nº Admitidos	% Admitidos
Masculino	600	350	58,3
Feminino	600	250	41,6

Suponha, que a Tabela 2.3 seja separada pelos departamentos das universidades, como mostra a Tabela 2.4

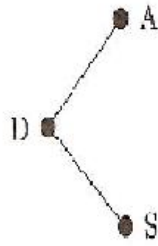
TABELA 2.4: Classificação cruzada por sexo, departamento e admissão

Departamento	Sexo	Nº Total	Nº Admitidos	% Admitidos
I	Masculino	100	25	25
	Feminino	300	75	25
II	Masculino	200	100	50
	Feminino	200	100	50
III	Masculino	300	225	75
	Feminino	100	75	75

Nos departamentos os homens têm proporção maior de admissão, mas não há evidência de discriminação sexual na admissão dentro dos departamentos. Da Tabela 2.3, vemos que Sexo (S) e Admissão (A) são marginalmente associados, em outras palavras, Admissão não é independente em relação à Sexo. A Tabela 2.4, indica que Admissão é independente de Sexo dado o Departamento (D), isto é, $S \perp\!\!\!\perp A | D$, como mostra a Figura 2.28

Neste exemplo, há associação marginal mas não associação condicional. Estritamente falando, o Paradoxo de Simpson refere-se a uma reversão na direção de associação entre a distribuição marginal e a condicional.

Exemplo 2.6. (*Ingresso em Universidade*): O exemplo final com dados discretos consiste na análise de uma tabela mostrando ingressos na universidade de Berkeley no outono de 1973, classificações - cruzadas pelo departamento e sexo do candidato (Freedman,

FIGURA 2.28: Grafo $S \perp\!\!\!\perp A | D$.

1978). Estes dados, mostrados na Tabela 2.5 são também analisados em Agresti (1990). Parece haver discriminação contra candidatas mulheres visto que apenas 24.94% dessas são admitidas contra 37.79% de homens.

TABELA 2.5: Admissões em Berkely no outono de 1973

Sexo	Admitidos	Não Admitidos	% Admitidos
Masculino	633	1042	37.79
Feminino	337	1015	24.92

Calculando a deviance dos dados usando software R, obtemos o valor de 54.22105 que nos sugere dependência entre Sexo e Admissão, ou seja, há evidência de discriminação sexual na universidade.

Esta conclusão não é correta, visto que é necessário levar em consideração os diferentes departamentos em que houve ingressos. Se rotularmos as variáveis admissão (A), departamento (D), e sexo (S) então a questão de interesse é aceitar (hipótese H_0), ou não, $A \perp\!\!\!\perp S | D$. Para examinar isto mais certamente, relembramos que $A \perp\!\!\!\perp S | D$ afirma que para todo departamento (D), ingresso (A) é independente de sexo (S).

Seguindo esse raciocínio, obtemos admissões dos quatro departamentos considerados na Tabela 2.6

TABELA 2.6: Admissões em Berkely no outono de 1973, considerando os departamentos

Departamentos	Sexo	Admitidos	Não Admitidos
I	Masculino	353	207
	Feminino	17	8
II	Masculino	120	205
	Feminino	202	391
III	Masculino	138	279
	Feminino	94	299
IV	Masculino	22	351
	Feminino	24	317

O teste de razão de verossimilhança de independência condicional dado o departamento, é dado por

$$d^2 = 2 \sum_{d=1}^4 \sum_{s=1}^2 \sum_{a=1}^2 n_{asd} \ln \left(\frac{n_{asd} n_{n_{++d}}}{n_{a+d} n_{+sd}} \right)$$

$$= \sum_{d=1}^4 \left\{ 2 \sum_{s=1}^2 \sum_{a=1}^2 n_{asd} \ln \left(\frac{n_{asd} n_{n_{++d}}}{n_{a+d} n_{+sd}} \right) \right\}$$

onde a expressão dentro de $\{.\}$ é o teste de deviance para independência no d-ésimo departamento.

Com subdivisão nos departamentos, calculamos suas respectivas deviances e p-valor apresentados na Tabela 2.7

TABELA 2.7: Admissões em Berkely no outono de 1973, considerando os departamentos

Departamentos	Deviances	G. L.	p-valor
I	0.2586429	1	0.6110540
II	0.7509844	1	0.3861648
III	0.9903864	1	0.3196480
IV	0.3836167	1	0.5356740

Observando as deviances, temos que Departamento e Admissão não são associados, ou seja, não há discriminação sexual, como afirmado anteriormente. Desta forma, vemos que os dados podem ser mostrados através de dois grafos, dados pela Figura 2.29

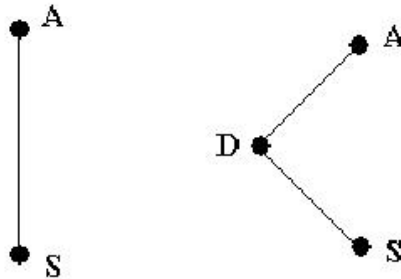


FIGURA 2.29: Primeiro Grafo: relação de departamentos conjuntos e segundo grafo: departamentos I, II, III, IV analisados separadamente.

Este é um exemplo do paradoxo de Simpson, onde vemos que há reversão no sentido da associação entre as variáveis Admissão e Sexo, quando controlamos o efeito da variável Departamento.

O modelo loglinear pode ser utilizado para identificar as relações de dependência entre variáveis discretas, apresentadas na forma de uma tabela de contingência. O grafo correspondente à relação entre variáveis discretas pode ser obtido a partir do modelo loglinear.

Além disso, o Paradoxo de Simpson evidencia o risco de análises simplistas em variáveis desta natureza.

Capítulo 3

Modelos Grafos Esparsos

3.1 Introdução

O interesse em métodos que exploram associações entre algumas variáveis tem sido fortalecido pelos problemas que envolvem estudo de expressão gênica de larga-escala.

Problemas de modelagem sob o paradigma “p grande, n pequeno” desafiam a estatística atual, tanto conceitual como computacionalmente.

Neste capítulo discutimos alguns aspectos do uso de modelos grafos Gaussianos de larga-escala na avaliação das associações em dados de expressão gênica.

Especificamente, Dobra *et al.* (2003) desenvolveram métodos para modelar a estrutura de covariância de distribuições de alta-dimensão com foco na estrutura *esparsa* - particularmente relevante na modelagem de associações em dados de expressão gênica. Desenvolveram também uma prova construtiva que generaliza os modelos grafos Gaussianos não orientados de grande escala, baseados na representação da distribuição conjunta de variáveis via conjunto de regressões lineares.

A relevância de formulações em termos de grafos acíclicos orientados e distribuições condicionais, redes dependentes são também discutidas e utilizadas.

A noção de esparsidade em modelos grafos para dados de expressão gênica reflete uma perspectiva, na qual um gene g , é bem predito por um subconjunto relativamente pequeno de genes. Essa formulação reflete o que é esperado em modelos grafos esparsos:

que cada variável ou gene tenha um número pequeno de vizinhos.

Em um modelo grafo, vizinhos de primeira-ordem um gene específico g são aqueles genes que, fazem g condicionalmente independente de todos os outros genes. Alguns genes identificados como vizinhos de g podem ser muito pouco correlacionados com g , mas muito relacionados com outros vizinhos em termos de correlação parcial.

Então, na exploração de relações transitivas entre pares ou conjuntos de genes baseados em grafos, o interesse nestes modelos encontra-se na capacidade de expandir a habilidade para identificar genes candidatos, que podem ser importantes numa trajetória específica sob investigação. Estes pontos são realçados no exemplo de câncer de mama que relaciona o receptor estrógeno com a proliferação e controle da célula Rb - E2F.

3.2 Grafos, Redes e Regressões

Estamos interessados em modelos de dependência em distribuição normal p - dimensional. Denotamos x um vetor normal p - variado. Assumimos dados centrados medidos em uma escala comum, e denotamos o modelo por $x \sim N_p(0, \Sigma)$ com matriz variância positiva definida e matriz precisão $\Omega = \Sigma^{-1}$. Denotamos os elementos de Ω por ω_{ij} , e os de Σ por σ_{ij} .

Nas aplicações x é um vetor de níveis de expressão gênica e p é muito grande; o exemplo tem $p = 12.558$. Escrevemos x_i para o i -ésimo elemento univariado de x , e para qualquer subconjunto de índices de $k \in \{1, 2, \dots, p\}$, escrevemos $x_k = \{x_i : i \in k\}$. Dois subconjuntos importantes são $[-i] = \{1, 2, \dots, p\} \setminus \{i\}$ e $k : m = \{k, k + 1, \dots, m\}$ para qualquer $k \leq m$. A distribuição conjunta implica o conjunto de distribuições univariadas condicionais $p(x_i | x_{[-i]})$ para $i \in \{1, 2, \dots, p\}$.

3.2.1 Modelos Grafos Gaussianos Esparsos

Escrevemos $ne(i) = \{j : \omega_{ij} \neq 0\}$, tal que as variáveis $x_{ne(i)}$ são os preditores de x_i - as variáveis que fazem x_i condicionalmente independente de todos os outros x_k para $k \notin ne(i)$. Então $p(x_i | x_{[-i]}) = p(x_i | x_{ne(i)})$. Na teoria de grafos e em termos de modelagem gráfica, $x_{ne(i)}$ define os vizinhos de x_i (ver Hofmann, Tresp, 1998). Nosso foco é em

modelos esparsos, com cada regressão condicional envolvendo um conjunto relativamente pequeno (para p) de preditores, tal que cada linha (portanto coluna) de Ω terá muitas entradas zero, então o número de $ne(i)$ será pequeno. Nisto vemos que, embora um gene possa ser associado com muitos outros genes, a “influência” desses genes será muitas vezes representada por um número pequeno de genes “preditivos”, diretamente em uma pequena vizinhança de x_i , consistente com a esparsidade de Ω .

Para fazer conexão com modelos grafos não orientados, escrevemos $V = \{1, 2, \dots, p\}$ e definimos o grafo $G = (V, E)$, onde V são os nós ou vértices, e E é o conjunto de arestas tal que, uma aresta existe entre os nós i e j se, e somente se, $j \in ne(i)$ (por simetria, se, e somente se, $i \in ne(j)$). A distribuição Normal $p(x)$ define um modelo grafo sob o grafo G (ver Lauritzen, 1996) - uma distribuição conjunta para x , na qual a estrutura de independência condicional é explicitamente exibida, através do conjunto de arestas caracterizando a $p(x_i|x_{ne(i)})$.

3.2.2 Redes de Dependência

O desafio de ajustar distribuições de alta dimensão pode ser focado por conjuntos ajustados de distribuições condicionais univariadas. Uma rede de dependência é simplesmente uma coleção de distribuições condicionais $\{p(x_i|x_{[-i]})\}$ que são definidas e construídas separadamente.

No contexto específico, de modelos normais esparsos, estes definem um conjunto de p regressões lineares condicionais separadas nas quais x_i é modelada por um subconjunto pequeno selecionado de outras variáveis, dizemos $x_{pv(i)}$ para algumas variáveis preditoras $pv(i)$, cada uma determinada separadamente.

3.2.3 Redes Composicionais

Uma forma direta para especificar um conjunto de modelos univariados de acordo com Dobra, Hans, Jones, Nevins e West (2003) é através de um modelo “triangular” que define a distribuição conjunta por composição, isto é, $p(x) = \{\prod_{i=1}^{p-1} p(x_i|x_{(i+1):p})\} p(x_p)$. Para isto é necessária uma ordenação específica de variáveis, as quais tomamos aqui como $1, 2, \dots, p$, sem perda de generalidade.

Suponha ajustar o conjunto de p regressões univariadas implicadas aqui, selecionando subconjuntos específicos de variáveis preditoras composicionais $\{cpv(i) \subseteq (i+1) : p\}$ para cada i (e com $cpv(p)$ vazio), e escolhendo parâmetros da regressão e valores da variância condicional para cada i . Isto define um conjunto de modelos

$$x_i = \sum_{j \in cpv(i)} \gamma_{ij} x_j + \epsilon_i \quad (3.1)$$

onde $\epsilon_i \sim N(0, \psi_i)$. A composição da distribuição conjunta $p(x)$ pode então ser escrita em forma estrutural como

$$x = \Gamma x + \epsilon, \epsilon \sim N_p(0, \Psi) \quad (3.2)$$

onde $\epsilon = (\epsilon_1, \dots, \epsilon_p)'$, $\Psi = \text{diag}(\psi_1, \dots, \psi_p)$, e Γ uma matriz triangular superior $p \times p$ com zero na diagonal.

Conectando com modelos grafos, definimos grafos acíclicos orientados (GAO) sob os nós V através das arestas E que são orientadas de cada elemento de $cpv(i)$ para i para todos $i \in V$. Por construção, este grafo acíclico, e a distribuição conjunta $p(x)$ definida pela composição representam uma rede Bayesiana sob o GAO (ver Lauritzen, 1996). Além disso, a estrutura corresponde a um grafo não orientado que é obtido diretamente pela substituição de todas as setas por arestas não orientadas, e “casando” todos os pares de pais com a ligação de uma aresta. As arestas resultantes correspondem a elementos não-zeros de Ω - os nós i, j estão ligados por uma aresta se, e somente se, $\Omega_{ij} \neq 0$. Este é um grafo não orientado - que define diretamente o conjunto de vizinhos de qualquer nó e exibe a estrutura de independência condicional da distribuição conjunta implicada.

Escrevemos $\theta_i = (\gamma_i, \psi_i)$ onde γ_i é o vetor coluna de coeficientes de regressão $\{\gamma_{ij} : j \in cpv(i)\}$ e o conjunto $\{\theta = \theta_i : i \in V\}$. Então, incluindo os parâmetros na notação, temos $p(x|\theta) = \prod_{i=1}^{p-1} p(x_i|x_{(i+1):p}, \theta_i)p(x_p|\theta_p)$ (notar que $\theta_p = \psi_p$).

A forma reduzida das redes Bayesianas 3.2 é imediata: $x = (I - \Gamma)^{-1}\epsilon$ tal que $\Omega = (I - \Gamma)^{-1}\psi_{-1}(I - \Gamma)$. Para qualquer conjunto de valores de parâmetros θ podemos então computar Ω e, por inversão, Σ .

3.3 Análise e Seleção de Regressões

Focaremos agora na regressão univariada para uma escolha de x_i , com k_i preditores em uma variável preditora indexada no conjunto $pv(i) \subseteq [-i]$, onde $(x_i|x_{pv(i)}, \theta_i) \sim N(x'_{pv(i)}\gamma_i, \psi_i)$ com $\theta_i = (\gamma_i, \psi_i)$. No caso geral o conjunto preditor pode envolver quaisquer das $p - 1$ variáveis $x_{[-i]}$. Nas seções seguintes, estes resultados se aplicam a escolha específica de $pv(i)$ que surgem nas redes de dependência (onde os elementos de $pv(i)$ podem ser quaisquer valores em $[-i]$) e então, nas redes composicionais (onde, dada a ordenação definida de variáveis, os elementos de $pv(i)$ são restritos a valores em, $(i+1) : p$).

Uma idéia estatística principal é a consistência de modelos e *prioris* com o número de variáveis preditoras, k_i , escolhidas. Estamos interessados em comparar modelos com diferentes candidatos de conjuntos preditores $pv(i)$, e cada modelo terá sua própria *priori* para os parâmetros θ_i implicados.

3.3.1 *Prioris* de Parâmetros de Regressão

Retornemos ao conjunto completo de p variáveis $x \sim N_p(0, \Sigma)$. Qualquer subconjunto de índices especificados $a(i) = (i, pv(i))$ definem o correspondente θ_i como uma transformação de elementos de Σ . A seleção de subconjuntos pode ser guiada por uma seleção de *prioris* de probabilidade de inclusão/exclusão para variáveis. No modelo abrangido, utilizamos a *priori* Wishart/ Wishart Inversa que é simples e efetiva em termos de características estatísticas. Começando com uma *priori* Wishart Inversa (Dawid, 1981; Dawid e Lauritzen, 1993), $\Sigma \sim IW_p(\delta, \tau I)$ com δ graus de liberdade e matriz escalar τI_p , para algum $\tau > 0$.

Da teoria normal padrão, a regressão é $x_i \sim N(x'_{pv(i)}\gamma_i, \psi_i)$ onde $\gamma_i = \Sigma_{pv(i)}^{-1}k_i$ e $\psi_i = \sigma_{ii} - k'_i\Sigma_{pv(i)}^{-1}k_i$. Usando a teoria da distribuição padrão deduzimos a *priori* implicada da gama normal/inversa da forma

$$(\gamma_i|\psi_i) \sim N_{k_i}(0, \tau^{-1}\psi_i I_{k_i}) \text{ e } \psi_i^{-1} \sim Ga((\delta + k_i)/2, \tau/2). \quad (3.3)$$

3.3.2 Verossimilhança Marginal e Distribuições *a Posteriori*

Consideremos agora a *posteriori* e aspectos preditivos de análises no processamento de n observações. Escrevemos y_i para o n -vetor de valores observados de x_i , e Z_i para a matriz $n \times k_i$ definida pelas correspondentes n observações no vetor predictor $x_{pv(i)}$. A forma do vetor é simplesmente $y_i \sim N_n(Z_i\gamma_i, \psi_i I_n)$ e *posterioris* e distribuições preditivas seguem o critério gama normal padrão/inversa da forma (ver Zellner, 1971)

$$(\gamma_i | \psi_i, y_i, Z_i) \sim N_{k_i}(M_i^{-1} Z_i' y_i, \psi_i M_i^{-1})$$

onde $M_i = \tau I_{k_i} + Z_i' Z_i$ e

$$(\psi_i^{-1} | y_i, Z_i) \sim Ga((\delta + k_i + n)/2, (\tau + q_i)/2)$$

onde $q_i = y_i' y_i - y_i' Z_i M_i^{-1} Z_i' y_i$.

Para comparar modelos baseados em subconjuntos diferentes de variáveis predictoras precisamos do valor do modelo de verossimilhança marginal, isto é, o valor observado da *priori* preditiva da função densidade $p(y_i | Z_i)$. Disto segue (ver Zellner, 1971 e Giudici, 1994) que

$$p(y_i | Z_i) \propto \frac{\tau^{k_i/2} \Gamma((n + \delta + k_i)/2)}{|M_i|^{1/2} \{1 + q_i/\tau\}^{(n+\delta+k_i)/2} \Gamma((\delta + k_i)/2)}. \quad (3.4)$$

Precisamos considerar o caso especial do modelo nulo - quando não há preditores. Formalmente, $k_i = 0$ e γ_i é nulo. O resultado será como acima, mas com $k_i = 0$ e $q_i = y_i' y_i$.

O interesse em modelos esparsos recomenda *prioris* e subconjuntos de variáveis $pv(i)$ na rede composicional que é penalizada pelos grandes valores de $k_i = \#pv(i)$. Uma *priori* tradicional para seleção de variáveis, especificada em termos de uma probabilidade comum de inclusão/ exclusão, é um primeiro passo, e a base da análise nos exemplos explorados aqui. Assumindo uma probabilidade comum de inclusão β , a probabilidade *a priori* de um modelo de um conjunto específico de k preditores relativos a um diferente, mas possivelmente sobreposto, conjunto de preditores $k + h$ específicos é apenas $(1 - \beta)^h \beta^{-k}$. Para ser efetivo em problemas com muitos preditores candidatos, a probabilidade de inclusão de variável tem que ser muito pequena. A *priori* binomial implicada no número de preditores em uma regressão pode ser usada para analisar isto. Com r preditores

candidatos, para β valores menores que $1/r$ a massa *a priori* de $k = 0$ aumenta, para um valor de, por exemplo, aproximadamente 0.6 se $\beta = 0.5/r$ e aproximadamente 0.9 com $\beta = 0.1/r$. Estes valores servem como referências rústicas para o que “pequeno” significa relativo ao número de preditores disponíveis. Valores de β menores penalizam a complexidade mais fortemente e induzem a esparsidade da rede composicional resultante.

Com *prioris* especificadas podemos procurar o modelo esparso estocasticamente ou deterministicamente. O exemplo aqui, é baseado na seleção de variáveis usando *forward/backward* para identificar um conjunto de regressões localmente ótimas, uma para cada x_i . Para cada i , aplicamos seleção *forward/backward* com respeito ao conjunto definido de candidatas preditoras, e contando cada regressão pelo modelo de probabilidades *a posteriori*, para uma regressão nos k parâmetros de um conjunto de candidatos r , a probabilidade *a posteriori* não normalizada é $\beta^k(1 - \beta)^{r-1-k}$ multiplicada pelo modelo de verossimilhança marginal da equação 3.4.

3.4 Construção de Modelos Grafos de Regressões

A abordagem baseada em Dobra *et al.* (2003) gerando modelos grafos esparsos é construída nas de redes de dependência e regressões univariadas. Em resumo:

- Construimos conjuntos de regressões em uma rede de dependência, e usamos essas para gerar uma ordenação apropriada de variáveis para ser a base de uma rede composicional;
- Com esta ordenação, a seleção de variáveis para cada modelo linear para variáveis ordenadas produz um GAO;
- Este GAO implica um único grafo não orientado, e distribuições *a posterioris* para os conjuntos de parâmetros de regressão no GAO então providenciam resultados para inferências nos grafos não - orientados.

Seja $\{(x_i|x_{pv(i)}) : i = 1, 2, \dots, p\}$ uma rede de dependência obtida maximizando a *posteriori* da regressão de probabilidades $p(x_i|x_{pv(i)})(\beta/(1 - \beta))^{\#pv(i)}$, $i = 1, 2, \dots, p$. A

probabilidade associada com a rede dependente é o produto

$$\prod_{i=1}^p p(x_i | x_{pv(i)}) (\beta / (1 - \beta))^{\#pv(i)}, \quad (3.5)$$

onde $\#pv(i)$ é o número de variáveis preditoras para cada variável i .

A rede dependente é uma rede composicional, então a equação 3.5 é proporcional à probabilidade *a posteriori* do modelo GAO implicado. Geiger e Heckerman provam que, sob *prioris* para parâmetros de regressão induzidos pela *priori* de Wishart inversa na matriz covariância Σ , quaisquer dois GAO's em algumas classes equivalentes (ver Anderson, Madigan, Perlman, 1997) têm alguma verossimilhança marginal. Entretanto, dois GAO's equivalentes têm um número de arestas orientadas e daí eles têm probabilidade *a priori* $(\beta / (1 - \beta))^\alpha$; portanto cada GAO tem probabilidade *posteriori* proporcional a 3.5.

Em uma rede dependente, os preditores para cada variável x_i podem ser selecionados de todas as outras variáveis. Uma rede composicional restringe esta escolha, preditores candidatos para x_i são restritos a variáveis que sucedem x_i na ordenação.

Como resultado, o *score* em 3.5 representa um limite superior para a probabilidade *a posteriori* não - normalizada do GAO resultante da rede composicional $\{(x_i | x_{cpv(i)}) : 1, 2, \dots, p\}$.

3.4.1 Construção de Modelo

Baseado em Dobra *et al.* (2003) temos

- Inicialização de redes de dependência

Para cada i , encontre um modelo de regressão $(x_i | x_{pv(i)})$ procurando sobre todas as variáveis $pv(i) \subseteq [-i]$. Utilize método de seleção *forward/backward* (por exemplo), para encontrar modelos *a posteriori* locais no espaço do modelo de regressão com conjuntos preditores $pv(i)$.

Coloque o contador de iteração $h = 0$. Inicialize um vetor ordenado de índice variável O como vetor nulo, e o índice variável candidato $C = V = \{1, 2, \dots, p\}$. Para cada $x_j, j \in C$, o *score* de explicação será

$$s_j^0 = \prod_{i \in C} p(x_i | x_{pv(i) \setminus \{j\}}) (\beta / (1 - \beta))^{\#pv(i)-1}.$$

Notamos que, quanto menor o *score*, maior é a “importância” desta variável na explicação das outras variáveis. O *score* de x_j pode ser facilmente obtido do *score* da atual rede de dependência $\{(x_i | x_{pv(i)}) : i \in C\}$ recalculando a probabilidade *a posteriori* da regressão em que x_j é uma variável explicativa.

- Ordenação iterativa de variáveis e geração de GAO’s composicionais

Para $h = 1, 2, \dots, p - 1$, repita o seguinte

1. Escolha a variável candidata com maior *score*, $j_h = \operatorname{argmax}_{j \in C} \{s_j^O\}$.
2. A variável x_{j_h} torna-se a próxima na ordenação boa, gravada por uma atualização do vetor ordenado de índice variável O a $[O, j_h]$. Grave $cpv(i) = pv(j_h)$ para esta variável.
3. Remova j_h do conjunto de índices de variáveis candidatas por uma atualização de C para $C \setminus \{j_h\}$
4. Para todo $i \in C$ tal que $j_h \in pv(i)$, reselectione e encontre o modelo de regressão para x_i escolhendo preditores potenciais apenas de C , para possivelmente gerar modelos revisados. Esses modelos então definem atualizações dos conjuntos $pv(i)$ e *scores* explicativos s_i^0 para todo $i \in C$.

Quando $h = p$, existe uma única variável remanescente em C que completa a lista ordenada O . O resultado é uma rede composicional definida por O e o conjunto de índices $cpv(i)$, por exemplo, uma rede Bayesiana sob o GAO resultante. Agora nós podemos extrair inferências sobre vários aspectos do modelo, incluindo estimativas dos coeficientes de regressão e, por transformação, estimativas Ω e portanto qualquer componente de Σ que são de interesse. Finalmente, removendo as setas e casando todos os pais de qualquer nó no grafo nós entregamos o grafo “moralizado” resultante - o grafo não orientado, com a probabilidade *a posteriori* não-normalizada associada.

3.4.2 Amostragem para Gerar Grafos Múltiplos

Esta construção utilizando métodos de seleção *forward/backward* para encontrar modelos de regressão para ordenação de variáveis, é determinística. Isto é chamada aproximação determinística “gananciosa” e, como tal, é naturalmente sensível a dados de perturbações que levariam a outros modelos finais.

Geralmente múltiplos modelos de regressão candidatos para cada variável nas redes compostas permite-nos avaliar múltiplos GAO's resultantes, daí grafos não-orientados, e comparar e contrastá-los, e isto pode facilmente ser empregado no passo de geração de GAO.

Uma extensão da aproximação acima começa considerar a ordenação múltipla das variáveis *prioris* para gerar GAO's, como segue. Selecionando a variável com maior *score*, podemos considerar várias possibilidades de ordenação usando um conjunto de variáveis com maior *score*, e/ou amostras das variáveis de acordo com algumas probabilidades. Especificamente, usamos uma aproximação que terá variáveis i de acordo com probabilidades proporcionais para $(s_i^O)^a$ para algum parâmetro $a > 0$. Baseados nos experimentos iniciais de Dobra *et al.* (2003) usamos $a = 25$.

3.5 Análise de Expressão Gênica em Câncer de Mama

Dobra, Hans, Jones, Nevins e West (2003) desenvolveram análises nos dados de expressão de 158 amostras de câncer de mama, no estudo de fenótipos moleculares para predição clínica (ver West *et al.* 2001; Huang, West, Nevins, 2003 e Huang *et al.* 2003). A análise ajusta o modelo do conjunto completo de 12.558 conjuntos próprios nos *microarrays* Affymetrix U95aV2 (ver Huang *et al.* 2003). A análise adota hiper-parâmetro $\beta = 1/(p - 1)$, uma *priori* consistente com a esparsidade e proliferação limitada de arestas nos grafos resultantes. A média *a priori* de $k_i = \#cp(i)$ é então 1, com probabilidade *priori* de 0.3 em cada um dos $k_i = 0, 1$. Usando o parâmetro $a = 25$, temos um total de 150 grafos são gerados.

Os dois exemplos da análise de dados de câncer de mama, demonstram a utilidade desses modelos na geração de informações biológicas. As figuras relacionadas ao grafo são

mostradas a seguir

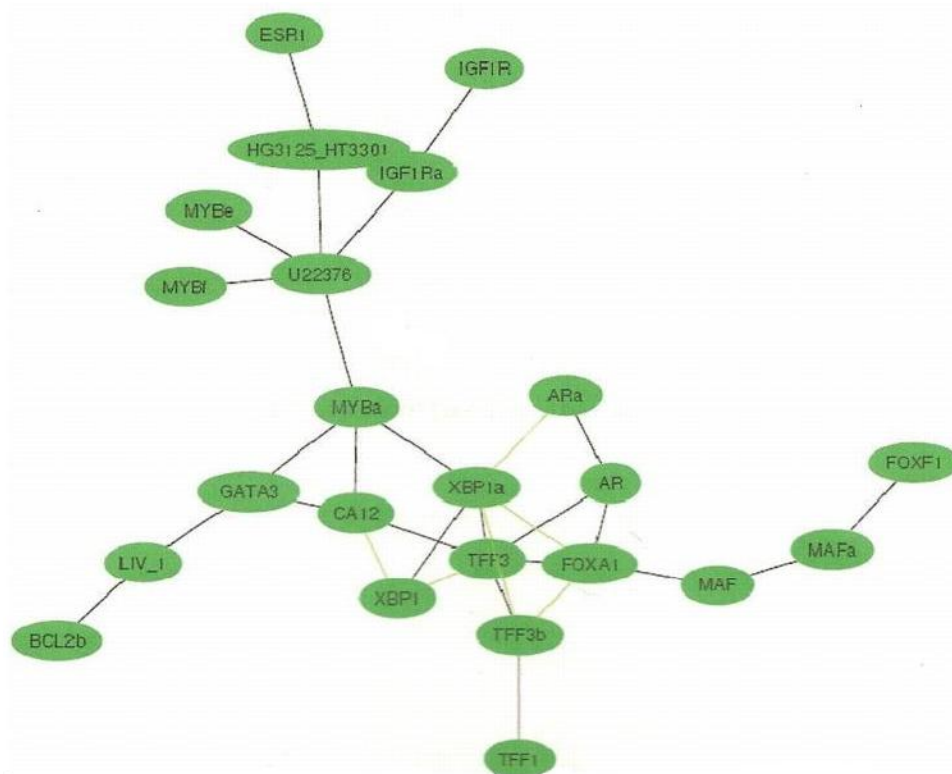


FIGURA 3.1: Subgrafos de câncer de mama baseados nos grafos da expressão gênica (em 12.558 genes) envolvendo ER e TFF1. Alguns genes tem múltiplos conjuntos de seqüências de oligonucleotídeos no *microarray*, e daí o aparecimento de alguns genes: estrógeno receptor (ESR1, HG3125 - HT3301), MYB (U22376, MYBa, MYBe, MYBf), AR (AR, ARa), c-MAF (MAF, MAFa), TFF3 (TFF3, TFF3b), XBP (XBP1, XBP1a) e IGF1R (IGF1R, IGF1Ra). O grafo mais provável gerado tem arestas indicadas em preto, as arestas adicionais verdes são aquelas aparecidas em pelo menos 5 dos 10 principais grafos gerados.

Fonte: Dobra, A., Hans, C., Jones, B., Nevins, J. R. e West, M. (2003)

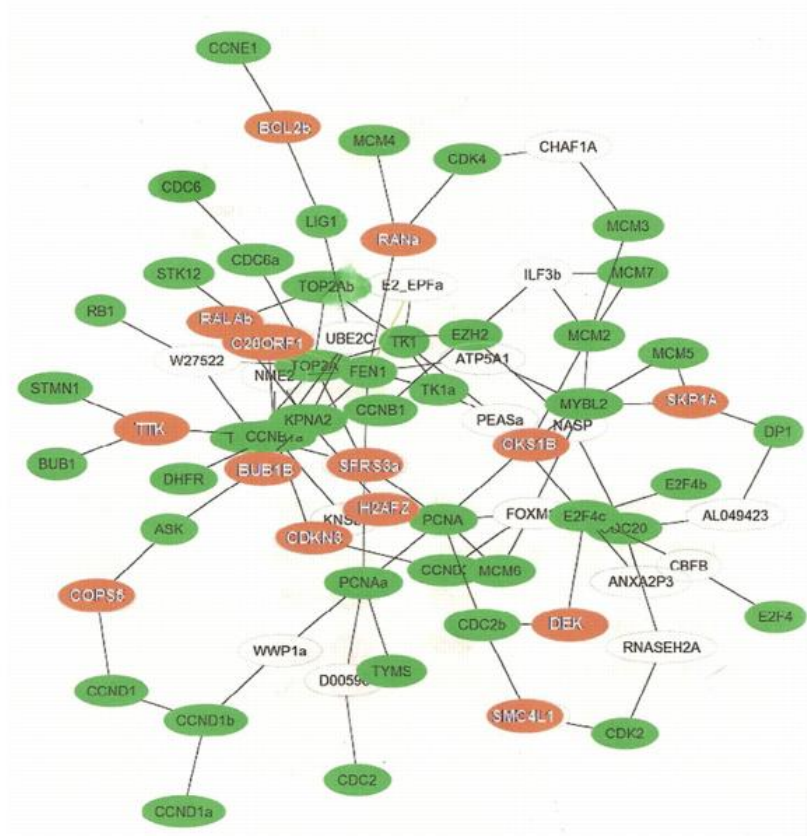


FIGURA 3.2: O subgrafo contendo Rb-E2F (verde) e ligados a genes (vermelhos e incolores), surgindo da exploração dos vizinhos de primeira-ordem do conjunto de genes em questão no *score* mais alto; 12.558 nós - genes na análise gráfica da expressão gênica dos dados de câncer de mama.

Fonte: Dobra, A., Hans, C., Jones, B., Nevins, J. R. e West, M. (2003)

3.6 Simulação

Baseados no trabalho de Dobra *et al.* (2003), onde foi feito um algoritmo para o desenvolvimento de modelos grafos esparsos, e também no trabalho de Seokho Lee (2006), desenvolvemos um programa com o objetivo de encontrar associações entre variáveis.

Utilizamos um conjunto de dados que relaciona as notas de 88 alunos em cinco matérias matemáticas, este conjunto de dados foi retirado de Whittaker (1990). Além desse conjunto de dados, simulamos um conjunto de dados normais.

Para o primeiro conjunto de dados, onde o resultado já era conhecido, conseguimos

obter o mesmo resultado, como podemos observar na Figura 3.3. Observamos que 1 corresponde à matéria de mecânica, 2 à vetores, 3 à álgebra, 4 à análise e 5 à estatística. A análise do relacionamento dessas matérias já foi feita na seção 2.4.

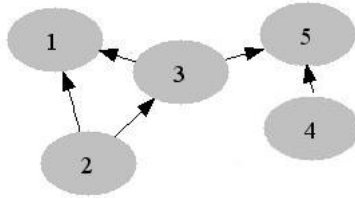


FIGURA 3.3: GAO que relaciona as cinco matérias matemáticas.

Este é um Grafo Acíclico Orientado (GAO), onde é necessário aplicar a teoria de moralização¹ onde trocam-se as setas por linhas e ligam-se as variáveis que têm os mesmos sucessores, 1 e 3 têm como mesmo sucessor 2, e 4 e 5 têm 3 como mesmo sucessor, desta forma obtemos o seguinte grafo acíclico não-orientado

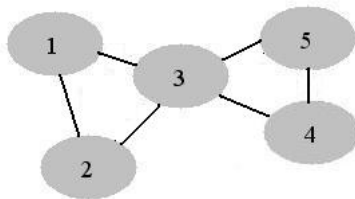


FIGURA 3.4: Grafo Acíclico não-orientado que relaciona as cinco matérias matemáticas.

O segundo conjunto de dados consiste de uma matriz de valores normais, com 10 linhas e 15 colunas, ou seja, 150 elementos. Para este conjunto de dados forçamos uma correlação entre as variáveis 3 e 4, 5 e 6, 7 e 8, 9 e 10. Esperamos que estas associações apareçam, observamos desta forma a Figura 3.5

¹Segue no Apêndice A a definição de moralização

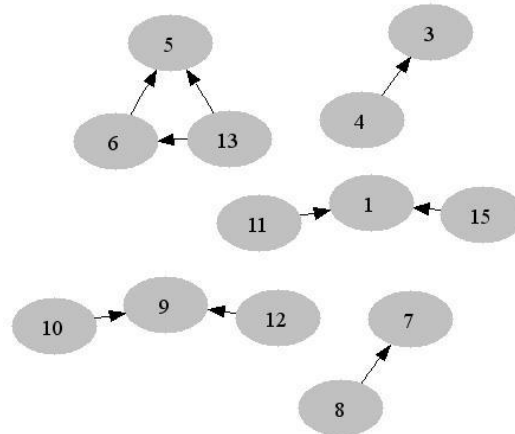


FIGURA 3.5: Grafo Acíclico Orientado que relaciona as 15 variáveis simuladas.

Fazendo o mesmo processo de moralização feito para o conjunto de dados anterior, obtivemos o seguinte grafo acíclico não-orientado

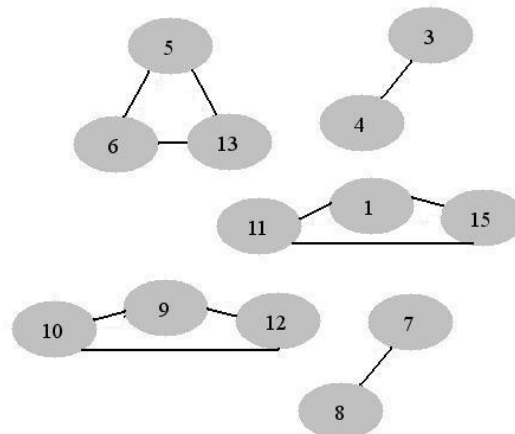


FIGURA 3.6: Grafo Acíclico não-orientado que relaciona as 10 variáveis simuladas.

Podemos observar que o grafo obtido, Figura 3.6, capta as associações esperadas.

Verificamos também, que para um conjunto de dados reais, o tempo consumido com 5963 variáveis/genes (Veer *et al.*, 2001) é enorme, tornando inviável a sua aplicação, seria necessário o uso de *cluster*, programação paralela, ou um computador muito eficiente, tais não estavam ao nosso alcance.

Mesmo sem essa aplicação, mas baseados em Dobra *et al.* (2003) e Seokho Lee (2006), constatamos que o algoritmo tem um bom funcionamento.

Capítulo 4

Seleção de Modelo para GAO's via RJMCMC

4.1 Introdução

O problema do espaço de busca aumentar exponencialmente com o número de variáveis incorporadas na análise confronta-se com a seleção de modelos. Por causa do enorme número de possíveis modelos, não é viável julgá-los por um critério de qualidade de ajuste como o AIC ou o BIC, portanto é interessante reduzir o espaço de busca, através de métodos como o RJMCMC.

Fronk (2002) propôs uma abordagem bayesiana usando *Reversible Jump* MCMC (RJMCMC) o qual foi introduzido por Green (1995) que é capaz de lidar com a variação de dimensão do espaço de busca atravessando este espaço de busca pela visita à vários modelos diferentes de acordo com suas probabilidades *a posteriori*. Além disso, este método é muito flexível e pode ser extensível para problemas nos quais ocorrem, por exemplo, dados *missing* ou variáveis latentes.

A seleção de modelos por RJMCMC tem sido desenvolvida por Giudici e Green (1999) para variáveis contínuas puras, e por Giudici, Green e Tarantola (1999) para o caso discreto puro. No contexto de GAO's um algoritmo de *Reversible Jump* para o caso contínuo tem sido introduzido por Fronk e Giudici (2000). Aqui baseados em Fronk (2002), apresentamos uma metodologia para GAO's Gaussianos.

4.2 Caso Gaussiano

Em Fronk e Giudici (2000), um algoritmo *Reversible Jump* para seleção de modelos em GAO's Gaussianos é proposto no qual as p variáveis aleatórias $x = (x_1, x_2, \dots, x_p)$ são assumidas seguindo uma distribuição normal multivariada. Então, para cada uma das distribuições condicionais univariadas $X_i|x_{pa(i)}$, $\beta_{i|pa(i)}$, $\sigma_{i|pa(i)}^2$ e d segue que

$$X_i|x_{pa(i)}, \beta_{i|pa(i)}, \sigma_{i|pa(i)}^2 \sim N \left(\beta_{i0} + \sum_{x_l \in pa(x_i)} \beta_{il}x_l, \sigma_{i|pa(i)}^2 \right)$$

onde $x_{pa(i)}$ denota o vetor conhecido das variáveis “pais” de X_i e $\beta_{i|pa(i)}$ o vetor dos coeficientes correspondentes. A variância condicionada é dada por $\sigma_{i|pa(i)}^2$ e d é o GAO fundamental. Além disso, as suposições seguintes são

$$\begin{aligned} \beta_{i|pa(i)}|\sigma_{i|pa(i)}^2, d &\sim N_{|pa(i)|+1} \left(b_{i|pa(i)}, \frac{1}{\alpha_i} \sigma_{i|pa(i)}^2 I \right), \\ \sigma_{i|pa(i)}^2|d &\sim IG(\delta_{i|pa(i)}, \lambda_{i|pa(i)}), \\ p(d) &= \frac{1}{a}, \end{aligned}$$

onde o vetor de coeficientes $\beta_{i|pa(i)}$ é também normalmente distribuído, a variância $\sigma_{i|pa(i)}^2$ segue uma distribuição gama inversa, e para o GAO d , uma distribuição uniforme discreta é assumida. O vetor de parâmetros $b_{i|pa(i)}$ e os parâmetros α_i , $\delta_{i|pa(i)}$ e $\lambda_{i|pa(i)}$ são escolhidos cuidadosamente, onde a denota o número de possíveis GAO's e I a matriz identidade. A seguinte fatoração da distribuição conjunta pode ser obtida:

$$\begin{aligned} p(x, \beta, \sigma^2, d) &= p(x|\beta, \sigma^2, d)p(\beta|\sigma^2, d)p(\sigma^2|d)p(d) \\ &= \prod_{i=1}^p p(x_i|x_{pa(i)}, \beta_{i|pa(i)}, \sigma_{i|pa(i)}^2) \prod_{i=1}^p (\beta_{i|pa(i)}|\sigma_{i|pa(i)}^2) \prod_{i=1}^p p(\sigma_{i|pa(i)}^2)p(d), \end{aligned}$$

onde $\beta = (\beta'_{1|pa}, \dots, \beta'_{p|pa})'$ e $\sigma^2 = (\sigma_{1|pa}^2, \dots, \sigma_{p|pa}^2)'$.

Para representar os GAO's e checar se são acíclicos usamos o conceito de matrizes de adjacência (ver Thulsirman e Swamy, 1992). Três diferentes passos devem ser levados em conta na busca de GAO's. Se dois vértices não são conectados, uma aresta pode ser inserida entre eles (passo “nascimento”), se existe aresta, esta pode ser deletada (passo

“morte”) ou trocar a orientação (passo “mudança”). A seguir temos um *loop* do algoritmo RJMCMC contendo 3 diferentes passos (ver Fronk, 2002).

4.2.1 Algoritmo *Reversible Jump* para o Caso Gaussiano:

1. Atualização de d por um passo de nascimento, morte ou mudança, para os passos de nascimento e mudança é necessário verificação da aciclividade do grafo.
2. Atualização de $\beta_{i|pa}$, $i = 1, \dots, p$.
3. Atualização de $\sigma_{i|pa}^2$, $i = 1, \dots, p$.

Notemos que o primeiro passo corresponde a seleção de modelo. Aqui, a troca na dimensão ocorre se um passo de morte ou nascimento é realizado. Os passos 2 e 3 atualizam os parâmetros de um GAO já existente, e portanto representa a estimação dos parâmetros.

Discutimos agora os diferentes passos em detalhes.

Passo de Nascimento e Morte:

De acordo com Fronk (2002), a Figura 4.1 mostra uma situação muito simples onde pode ser visto o “nascimento” de uma aresta. Ao adicionar uma aresta em um grafo real, isto pode resultar na adição de vários passos no grafo de interação. Em geral, pode ser declarado que adicionando uma aresta i a j no grafo real d implica $1 + |pa_m(j)|$ novas arestas no grafo de interação, onde $|pa_m(j)|$ denota o número de “pais” de j no grafo real. Conseqüentemente, o i -ésimo modelo de regressão é ampliado por $1 + |pa_m(j)|$ novas covariáveis.

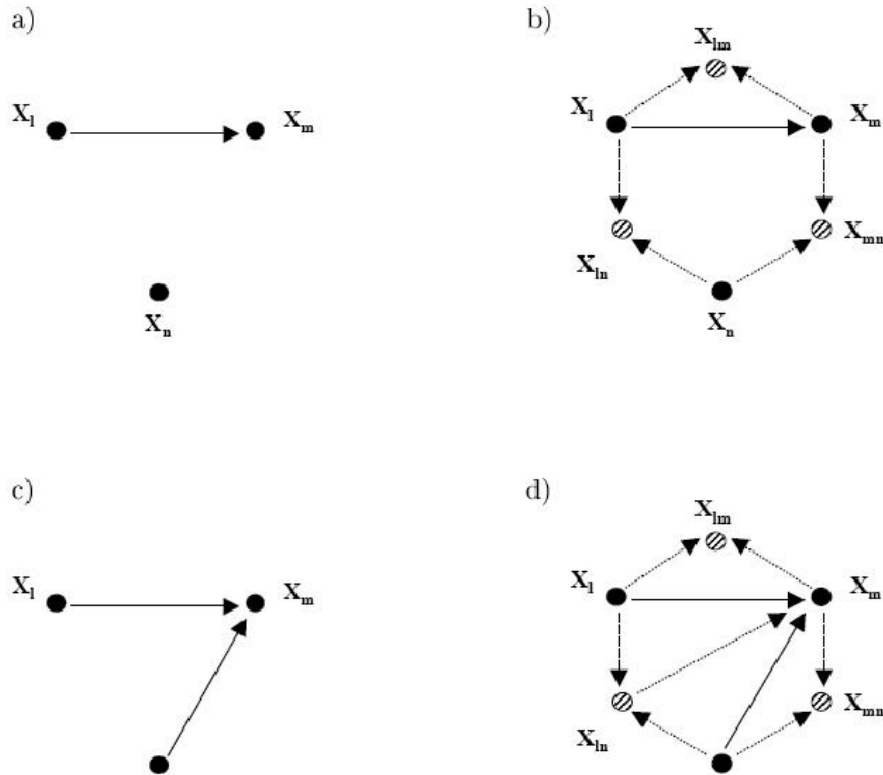


FIGURA 4.1: Exemplo para um passo de nascimento em um grafo com 3 vértices. O grafo real com 1 e 2 arestas é dado em (a) e (c), as interações correspondentes no grafo, são mostradas em (b) e (d).

Fonte: Fronk (2002)

Passo de Mudança:

Como é mostrado na Figura 4.2, mudando a aresta (j, i) para (i, j) , o número de coeficientes de regressão envolvidos nos dois modelos de regressão de i e j não permaneceram inalterados. Como consequência, trabalhando com o grafo de interação o passo de mudança pode também trocar a dimensão como os passos de nascimento e morte. Isto sempre ocorre se $|pa_m(i)| - 1 \neq |pa_m(j)|$. Em seguida, assumimos que o número total de parâmetros é aumentado, e chamamos passo de mudança 2, que reflete uma situação onde o número total de parâmetros é decrescente. As investigações de Fronk (2002) têm mostrado que uma boa probabilidade de aceitação é somente obtida quando todos os parâmetros dos modelos considerados são reamostrados.

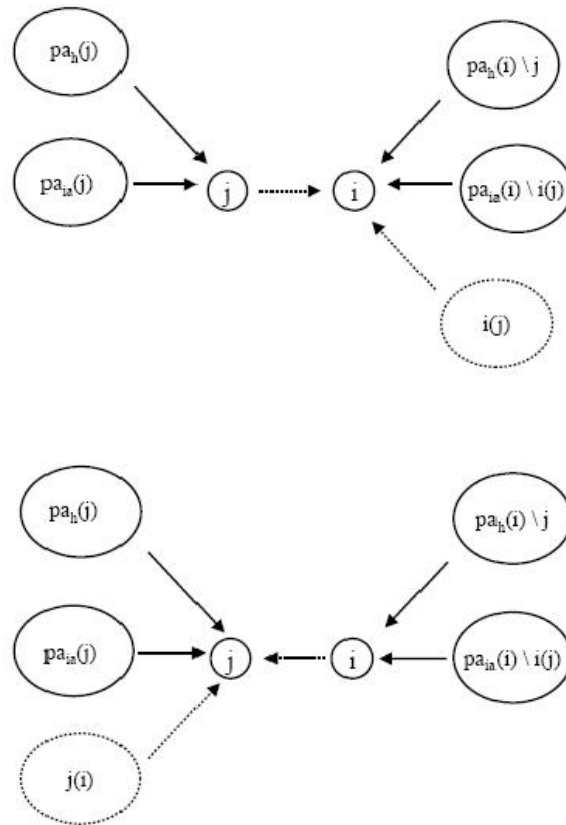


FIGURA 4.2: Troca no subgrafo envolvido do grafo de interação que é causado por um passo de mudança. Em ambos os casos, a tarefa dos pais refere-se ao grafo acima. Os conjuntos de pais de i e j não precisam ser disjuntos.

Fonte: Fronk (2002)

4.3 Exemplo: Mulher e Matemática

O conjunto de dados de interesse aqui tornou-se bastante famoso sob o nome “Mulher e Matemática”, origina-se de uma pesquisa de Lampagne (1979). O objetivo foi analisar o sucesso de uma palestra chamada “mulher e matemática”. Para esta finalidade, com 1190 estudantes de oito escolas, foram analisadas algumas variáveis demográficas como atitude para com a matemática e se elas tinham tomado conhecimento do programa. A Tabela 4.1 mostra as seis variáveis investigadas. Os dados tem sido analisados por vários autores como por Upton (1991), Madigan e Raftery (1994) e Giudici (1999).

TABELA 4.1: Variáveis do conjunto de dados “mulher e matemática”

Var.	Questão	Resposta
X_0	Comparecimento a aula	sim/não
X_1	Sexo	feminino/masculino
X_2	Tipo de Escola	urbano/suburbano
X_3	“Eu preciso de matemática para meu futuro”	concordo/não concordo
X_4	Preferência Subjetiva	matemática e ciência/artes liberais
X_5	Planos Futuros	estudo/trabalho

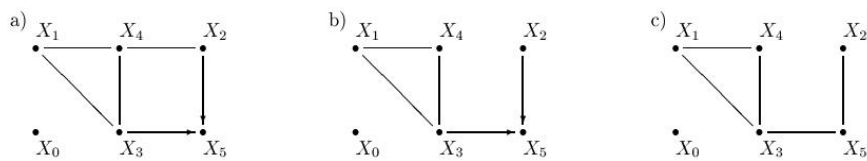


FIGURA 4.3: Os grafos essenciais com as maiores probabilidades *a posteriori* de 0.20 em (a) e 0.10 em (b) pelo algoritmo *Reversible Jump* para GAO's e grafo decomponível não orientado com a maior probabilidade (0.52) pelo algoritmo *Reversible Jump* hierárquico de Giudici *et al.* em (c).

Fonte: Fronk (2002)

Os resultados de Fronk (2002) são baseados em um *runtime* de 205000 iterações onde as primeiras 5000 são tratadas como *burn-in*. Além disso, usa-se o passo de mudança com aceitação de 0.5. Para resumir os GAO's diferentes de uma classe equivalente, apresentamos ela por seu grafo essencial. A Figura 4.3 mostra os dois modelos com as probabilidades *a posteriori* mais altas de 0.2 e 0.1. O melhor modelo de Giudici (1999), que é procurado pelo espaço de grafos decomponíveis não orientados, parece quase similar (ver Figura 4.3) e é selecionado com uma probabilidade de 0.52. As variáveis X_1, X_3 e X_4 , formam um clique. A principal diferença entre os resultados das duas abordagens em relação a do *Reversible Jump*, é que, essa estratégia suporta uma independência de X_2 e X_3 dado X_4 , onde desaparece o condicionamento em X_5 .

A imoralidade $X_3 \rightarrow X_5 \leftarrow X_2$, pode também ser vista na matriz de adjacência e esboço, onde as arestas do grafo essencial da Figura 4.3 são impressas em negrito. Chamamos S a matriz esboço que pertence a matriz de adjacência A se, $[S]_{ij} = [A]_{ij} + [A]_{ji}$

para $i, j = 1, \dots, p$, essa matriz dá informação sobre a estrutura geral do grafo mas ignora a orientação das arestas.

$$\bar{A}_{mem} = \begin{pmatrix} 0 & .03 & .03 & .03 & .03 & 0.1 \\ .04 & 0 & .02 & .4 & .36 & 0.00 \\ .05 & .02 & 0 & .06 & .13 & .77 \\ .07 & .60 & .22 & 0 & .45 & .82 \\ .05 & .64 & .33 & .55 & 0 & .21 \\ .07 & .00 & .23 & .18 & .11 & 0 \end{pmatrix}$$

$$\bar{S}_{mem} = \begin{pmatrix} * & .06 & 0 & .09 & .08 & .09 \\ * & * & .04 & \mathbf{1} & \mathbf{1} & .01 \\ * & * & * & .27 & .46 & \mathbf{1} \\ * & * & * & * & \mathbf{1} & \mathbf{1} \\ * & * & * & * & * & .32 \\ * & * & * & * & * & * \end{pmatrix}$$

A estrutura de associação torna-se quase clara pela matriz esboço (ver matrizes acima) onde aqueles vértices que são conectados nas Figuras 4.3b e 4.3c são ligados por uma aresta com probabilidade 1.

4.4 Conclusão

Fronk (2002), introduz um algoritmo *Reversible Jump* para seleção de GAO's para o caso contínuo. Estudos de simulação têm sido realizados para analisar a performance do algoritmo. Os resultados tornam-se satisfatórios. O exemplo do conjunto de dados “mulher e matemática” mostra que também para dados sensíveis reais são obtidos resultados.

Há muitas possibilidades de extensão deste algoritmo. O *design* flexível do MCMC oferece a possibilidade de também ser usado para variáveis latentes ou *missings*. O algoritmo tem sido considerado por Giudici e Stanghellini (1999) para grafos não orientados. Em muitos casos práticos, algumas das possíveis influências e dependências já

são conhecidas. Portanto, seria conveniente que a aresta fosse fixada durante a busca do modelo que, também implica na redução no tempo de execução, desde que o espaço de busca decresça enormemente.

Capítulo 5

Computação Estocástica para Modelos Grafos de Alta Dimensão

5.1 Introdução

A década de 90 testemunhou uma revolução essencial na ciência estatística, baseada no desenvolvimento de métodos de simulação estocástica para computação científica. O impacto na aplicação da estatística Bayesiana tem sido particularmente notável, com desenvolvimento de métodos MCMC, tornando possível a aplicação de modelos matemáticos cada vez mais valiosos e mais relevantes. Junto com esses modelos complexos está a capacidade de gerar conjuntos de dados envolvendo muitas variáveis. Das finanças de alta-freqüência e enormes conjuntos de dados de *marketing* e estudo de expressão gênica relacionado ao genoma, estamos agora frente a problemas que envolvem muitas variáveis. O uso de métodos de computação estocástica na procura de modelos candidatos aumenta o desafio e eficiência da estatística computacional.

Em um modelo grafo de uma distribuição multivariada, os nós representam variáveis e as arestas representam a dependência entre os pares de variáveis, tal que o conjunto de arestas define a estrutura de independência condicional global da distribuição. As questões metodológicas a frente, incluem, por exemplo, a especificação da distribuição *priori*. Estas questões representam um complexo problema de seleção de variável/modelo para a qual vários métodos computacionais têm sido sugeridos.

Focamos nos modelos grafos não-orientados, os quais têm gerado muitos interesses na estatística computacional recentemente. Um modelo grafo Gaussiano não-orientado está, em média (com respeito a distribuição *posteriori* da matriz covariância) sobre distribuições normais multivariadas, com uma estrutura de independência condicional comum. Estamos interessados em explorar “espaços grafos” e grafos identificados (estrutura de independência condicional) que estão mais apropriados para um determinado conjunto de dados quando medido pela distribuição *a posteriori* sobre os grafos.

Artigos recentes envolvem questões de melhoramentos dos métodos computacionais para estimação de modelos grafos Gaussianos não-orientados. Um foco fundamental nesta literatura tem sido a decomposição de modelos grafos. Giudici e Green (1999) mostram condições computacionais fáceis e eficientes para determinar se, ao adicionarmos uma aresta a um grafo, mantemos a propriedade de decomposição.

Esta linha geral de desenvolvimento tem sido estendida aos grafos decomponíveis, além dos métodos de Monte Carlo para calcular verossimilhanças marginais para modelos não-decomponíveis, Roverato (2002), Atay - Kayis e Massam (2005) e Dellaportas, Giudici e Roberts (2003) são de utilidade imediata. Alternativamente, métodos de aproximações e especificações da distribuição *priori* que substancialmente simplificam a computação dessas verossimilhanças marginais são desenvolvidas por Wong, Carter e Kohn (2003).

Apesar desses avanços e a necessidade de desenvolver mais metodologias para problemas de alta-dimensão, a literatura recente mantém um foco geral em problemas pequenos. Wong, Carter e Kohn (2003) são uma exceção.

Concluimos o capítulo revendo a especificação da distribuição *priori* e abordagens computacionais em modelos grafos Gaussianos com abordagens construtivas alternativas que são capazes de lidar com dimensões muito grandes, e comentamos também a necessidade para que cada vez mais, sejam desenvolvidas ferramentas computacionais que solucionem esse tipo de problema.

5.2 Modelos Grafos e Estrutura Gráfica

Modelos grafos providenciam representações da estrutura de independência condicional de uma distribuição multivariada, tão bem como o acesso a algoritmos eficientes para computação das densidades condicionais e marginais (ver Whittaker, 1990; Lauritzen, 1996; Anderson, Madigan, Perlman, Richardson, 1999; Cowell, Dawid, Lauritzen, Spiegelhalter, 1999). A eficiência computacional surge da decomposição dos espaços amostrais em subconjuntos de variáveis (vértices nos grafos) baseada nas suas relações gráficas. A distribuição conjunta dessas variáveis sobre grafos, como verossimilhanças e densidades *a priori* e *a posteriori*, podem ser calculadas separadamente nos subconjuntos de vértices e então juntadas novamente dentro de uma verossimilhança ou densidade incorporando todas as variáveis (ver Hammersley e Clifford, 1971; Dawid e Lauritzen, 1993).

A terminologia básica e idéias para modelos grafos (ver Cowell, Dawid, Lauritzen e Spiegelhalter, 1999) e a notação usada aqui, começa com um grafo $G = \{V, E\}$, onde G é definido sobre um conjunto de vértices V (as variáveis), e um conjunto de arestas E . Duas variáveis a e b são vizinhas em G se, e somente se, a aresta $(a, b) \in E$. Um grafo completo com p vértices tem todas as $\frac{p!}{2!(p-2)!}$ arestas, senão o grafo é incompleto. O grafo incompleto G decompõe-se em subgrafos disjuntos A , B e C (com $A \cup B \cup C = G$). Se C é completo e separa A e B (qualquer caminho de um vértice em A para um vértice em B passa através de C), o subgrafo C é um separador. A decomposição é própria se nem A nem B estão vazios. Se o separador C escolhido é minimal, então de forma iterativa, a decomposição nos subgrafos $A \cup C$ e $B \cup C$ resulta em componentes primos de um grafo - uma coleção seqüencialmente definida de subgrafos que não podem ser novamente decompostos. Ver Figura 5.1.

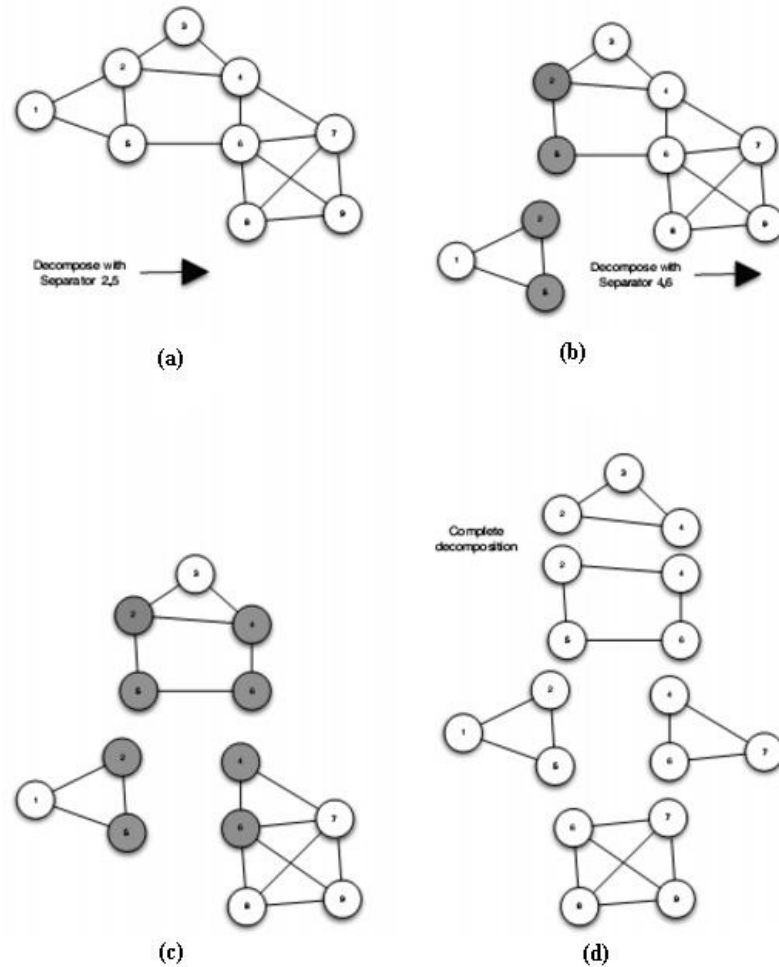


FIGURA 5.1: Exemplo de como decomposição iterativa de um grafo produz componentes primos.

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

Qualquer grafo conectado pode ser representado como uma árvore destes componentes primos - uma *árvore de junções*. Na árvore de junções, cada componente primo, denotado por P_i , é um nó; se dois nós compartilham de um conjunto de vértices, todos os componentes primos no caminho entre eles na árvore de junções também contêm o conjunto de vértices. Um conjunto de vértices compartilhados entre dois componentes primos formam um subgrafo completo (da definição de decomposição). Os conjuntos de vértices compartilhados pelos nós adjacentes na árvore de junções são chamados de separadores de árvores de junções, denotados por S_i . Um exemplo é mostrado na Figura 5.2.

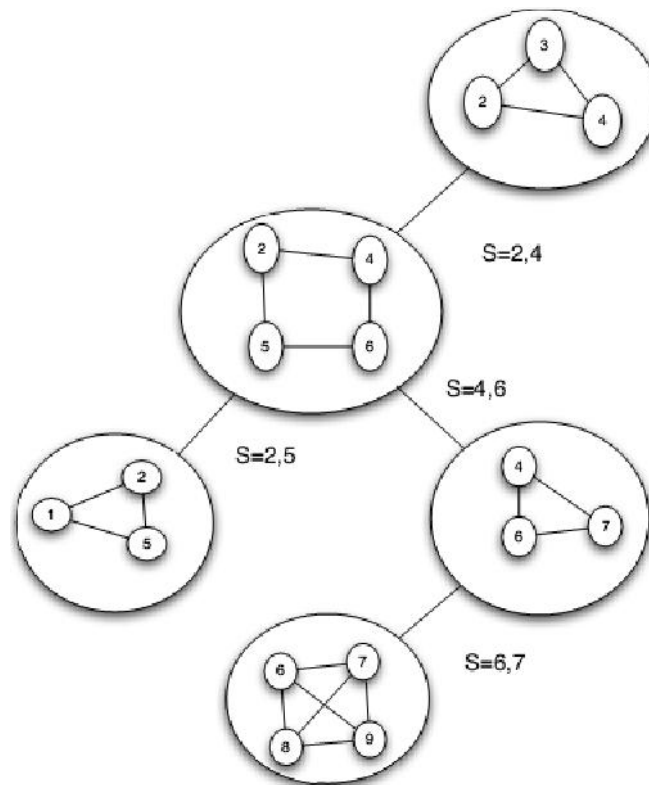


FIGURA 5.2: Representação de um grafo como uma árvore de junções, uma árvore de componentes primos.

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

Para grafos com mais que um componente conectado, uma árvore de junções existe para cada componente conectado; a coleção de árvores de junções é chamada *floresta de junções*. Quando uma floresta de junções não contem ciclos (*loops*) entre os nós, podemos definir uma *perfeita ordenação* de componentes primos e separadores. Denotamos uma ordenação de componentes primos (P_i) e separadores (S_i) como $P_1; S_2, P_2; S_3; P_3; \dots$ onde S_i é a intersecção de P_i e todos os componentes inferiores numerados. Chamamos a seqüência de componentes primos G^i e a seqüência de separadores, S^i . Notamos que mais que uma perfeita ordenação pode existir para qualquer grafo.

Se todos os componentes primos de um grafo são completos, o grafo é dito *decomponível*. Subgrafos completos maximais são chamados *cliques*, assim os componentes primos de um grafo decomponível são todos os cliques. Quando estivermos referindo exclusivamente a componentes primos que são cliques, usaremos C para denotar o componente ao contrário de P . Em modelos grafos Gaussianos, as propriedades das distribuições de

conjuntos de variáveis representados pelos grafos completos estão bem entendidas: elas têm distribuições normais multivariadas.

5.3 Modelos Grafos Gaussianos

5.3.1 Fatoração de Densidade e Verossimilhança

A fatoração de distribuições conjuntas que satisfazem a independência condicional pela estrutura da aresta de um dado grafo, é fundamental para desenvolver análises de modelos grafos. Em geral, um vetor aleatório p -dimensional da variável y tem uma distribuição multivariada $p(y)$ que, no grafo G especificado, fatora-se em termos correspondentes aos componentes primos e separadores de qualquer representação de árvores de junções de G , isto é,

$$p(y) = \frac{\prod_{P \in G^i} p(y_P)}{\prod_{S \in S^i} p(y_S)}, \quad (5.1)$$

onde y_P e y_S representam os subconjuntos de variáveis nos componentes primos e separadores, respectivamente. No caso especial de uma distribuição Gaussiana multivariada, a formulação em termos de uma matriz de covariância estruturada ou termos da covariância inversa, ou matriz precisão, claramente isola a estrutura chave.

Com uma matriz Σ de covariância não-singular, portanto a matriz de precisão $\Omega = \Sigma^{-1}$, cada termo na decomposição é uma distribuição Gaussiana multivariada com matrizes de covariância Σ_{PP} , Σ_{SS} nos componentes primos e separadores. Então, para uma amostra aleatória de tamanho n , $Y = (y_1, y_2, \dots, y_n)$, a função densidade conjunta no grafo G tem a representação:

$$p(Y|\Sigma_G) = \frac{\prod_{P \in G^i} p(Y_P|\Sigma_{PP})}{\prod_{S \in S^i} p(Y_S|\Sigma_{SS})}. \quad (5.2)$$

Isto fornece a função de verossimilhança completa para Σ em problemas de inferência para um dado grafo G . Dempster (1972) refere-se ao problema de identificação de estruturas gráficas através de zeros na matriz precisão, o uso desta terminologia pelos autores subsequentes incluem Wong, Carter, e Kohn (2003). Esta abordagem está caracterizada

no Ω associado a G através do conjunto de arestas E , no qual $\Omega_{ij} \neq 0$ se, e somente se, a aresta $(i, j) \in E$. A inferência formal é estruturada pela composição; de uma perspectiva Bayesiana, estamos interessados nas distribuições *a posteriori* $p(G, \Sigma|Y) = p(\Sigma|G, Y)p(G|Y)$ para distribuições *prioris* especificadas $p(G, \Sigma) = p(\Sigma|G)p(G)$.

5.3.2 *Prioris e Posterioris* para Matrizes de Covariância

Giudici (1996) discute a especificação da distribuição *priori* para Σ , comparando as “*prioris* locais” descritas em Dawid e Lauritzen (1993), e as “*prioris* globais” baseadas na abordagem condicional em Dickey (1971). Estas distribuições *prioris* baseadas na Wishart inversa têm a propriedade que $p(\Sigma|G)$ é consistente sobre grafos no sentido de manter uma distribuição *a priori* comum para o elemento (i, j) de Ω quando o grafo não contem o elemento (i, j) sendo zero. Giudici e outros autores previamente declararam as distribuições *prioris* locais, como sendo convenientes somente para modelos decomponíveis. Roverato (2002) estendeu esta classe de distribuições *prioris* também para modelos não-decomponíveis. Giudici (1996) sugere que *prioris* locais estimulam grafos esparsos, por esta razão usaremos distribuições *prioris* locais.

A distribuição *priori* $p(\Sigma|G)$ é uma Wishart hiper-inversa, $WHI(G, \delta, \Phi)$. Aqui Φ é uma matriz positiva definida e $\delta > 0$, são parâmetros definidos. Em modelos decomponíveis:

$$p(\Sigma|G) = \frac{\prod_{C \in G^i} p(\Sigma_{CC}|G)}{\prod_{S \in S^i} p(\Sigma_{SS}|G)} \quad (5.3)$$

Para cada componente primo completo C de G e cada separador, a correspondente submatriz da covariância, Σ_{CC} , tem uma distribuição *priori* Wishart Inversa (δ, Φ_{CC})

$$p(\Sigma_{CC}|G) = \frac{|\frac{\Phi_{CC}}{2}|^{\frac{(\delta+|C|-1)}{2}}}{\Gamma_{|C|}\left(\frac{\delta+|C|-1}{2}\right)} |\Sigma_{CC}|^{-\frac{\delta+2|C|}{2}} \exp\left\{\frac{-1}{2} \text{tr}(\Phi_{CC}\Sigma_{CC}^{-1})\right\}, \quad (5.4)$$

onde $\Gamma_k(a)$ é uma função gama multivariada,

$$\Gamma_k(a) = \pi^{\frac{k(k-1)}{4}} \prod_{i=0}^{k-1} \Gamma\left(a - \frac{i}{2}\right).$$

Grafos decomponíveis consistem completamente de componentes primos completos, as equações 5.3 e 5.4 são suficientes para expressar a densidade de Σ quando restringimos a grafos decomponíveis.

Grone *et al.* (1984) mostraram que quando considerando uma matriz de covariâncias incompleta (onde somente as entradas correspondendo a arestas ou a diagonal estão preenchidas), a matriz pode ser completada sendo um matriz positiva definida consistente com o grafo, este completamento é único. Grone também mostrou que as submatrizes correspondendo aos cliques em um grafo decomponível são positivas definidas, então uma completção positiva definida consistente com o grafo sempre existe. Isto é refletido na densidade para grafos decomponíveis: nenhum dos elementos “não-livres” aparecem em 5.3 ou 5.4, assim eles não afetam a densidade.

Em relação a grafos não-decomponíveis precisamos de uma expressão análoga a 5.4 para um componente primo não-completo, Roverato (2002) providencia uma generalização da Wishart inversa como uma densidade *a priori* para Σ_{PP} (então chamaremos a distribuição *priori* sobre Σ uma distribuição Wishart hiper-inversa como no caso decomponível). A distribuição *priori* é derivada como a conjugada Diaconis-Ylvisaker (ver Diaconis e Ylvisaker, 1979) da verossimilhança para Ω . Nesta densidade, alguns dos elementos não-livres de Σ_{PP} aparecem; entretanto, a verdadeira dimensão da densidade corresponde ao número de elementos livres. Os elementos livres são determinados pelo conjunto de arestas e , assim mostramos a densidade como Σ_{PP}^E . A expressão para a densidade *a priori* é então

$$p(\Sigma_{PP}^E|G) \propto |\Sigma_{PP}|^{-\frac{\delta-2}{2}} J(\Sigma_{PP}^E) \exp\left\{\frac{-1}{2}\Sigma_{PP}^{-1}\Phi_{PP}\right\}, \quad (5.5)$$

onde Σ_{PP} é a complementação da matriz positiva definida de Σ^E e $J(\Sigma_{PP}^E)$ é o Jacobiano da transformação de Ω_{PP}^E (que tem zeros para as entradas não-diagonais não correspondendo as arestas em E) para Σ_{PP}^E . Esta densidade é obtida de uma distribuição *priori* de Wishart em Ω_{PP} , condicionada em Ω_{PP} consistente com G , por uma mudança de variáveis. Baseado no condicionamento, esta distribuição *priori* difere da distribuição *priori* global de Giudici (1996) onde o condicionamento é somente usado dentro dos componentes primos. Quando usando a distribuição *priori* Wishart hiper-inversa com um espaço grafo não restrito (modelos não decomponíveis), Jones *et al.* (2004) garante

que δ é estritamente maior que 2.0.

A distribuição *priori* de Wishart hiper-inversa é conjugada em qualquer dos casos decomponível ou não restrito; a distribuição *posteriori* é Wishart hiper-inversa ($G, \delta^* = \delta + n, \Phi^* = \Phi + S_y$), onde S_y é a soma dos produtos da matriz, $\sum_{i=0}^n y_i y_i'$. Nos exemplos abaixo, usamos $\Phi = \tau I$ para especificar constantes τ (outras escolhas para Φ são consideradas em Giudici e Green, 1999). A forma da distribuição *posteriori* tornar-se limpa que é importante para a escolha de τ estar na escala apropriada, ou poder dominar o efeito dos dados. De fato, τ crescente promove probabilidade *a priori* crescente ou grafos mais complicados com mais arestas, como na Figura 5.3. O modelo *a priori* marginal para cada termo da variância (σ_{ii}) é $\tau/(\delta + 1)$; se os dados são padronizados, todas as variâncias serão 1.0 e τ poderá ser o conjunto $\delta + 1$ (ver Jones *et al.* 2004).

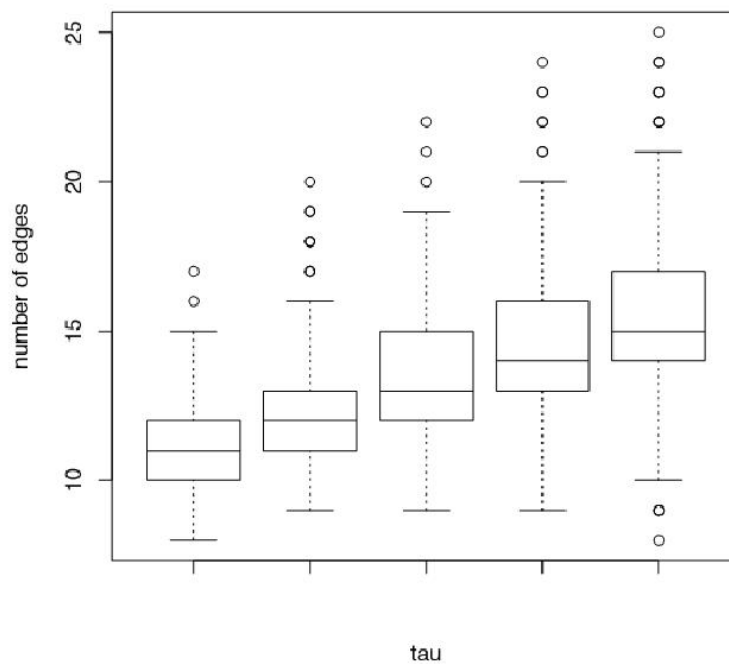


FIGURA 5.3: Boxplot das amostras *a posteriori* do número de arestas, para diferentes valores de τ .

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

5.3.3 Funções de Verossimilhança Marginal sobre Grafos

A função de verossimilhança marginal avaliada por qualquer grafo G é

$$p(Y|G) = \int_{\Sigma|G} p(Y|\Sigma)p(\Sigma|G)d\Sigma.$$

Notando que a constante normalizadora *a priori* e um fator de $(2\pi)^{-np/2}$ da verossimilhança pode ser colocado fora da integral, esta expressão torna-se uma função simples de constantes normalizadoras da distribuição *priori* e da distribuição *posteriori*, $h(G, \delta, \Phi)$ e $h(H, \delta^*, \Phi^*)$:

$$p(Y|G) = (2\pi)^{-np/2} \frac{h(G, \delta, \Phi)}{h(G, \delta^*, \Phi^*)}. \quad (5.6)$$

Para um grafo decomponível, as constantes normalizadoras da Wishart hiper-inversa são funções de constantes normalizadoras para as densidades do clique da Inversa de Wishart e separador, dados em 5.4

$$h(G, \delta, \Phi) = \frac{\prod_{C \in G_i} \left| \frac{\Phi_{CC}}{2} \right|^{\frac{(\delta+|C|-1)}{2}} \Gamma_{|C|} \left(\frac{\delta+|C|-1}{2} \right)^{-1}}{\prod_{S \in S_i} \left| \frac{\Phi_{SS}}{2} \right|^{\frac{(\delta+|S|-1)}{2}} \Gamma_{|S|} \left(\frac{\delta+|S|-1}{2} \right)^{-1}}. \quad (5.7)$$

Para grafos não decomponíveis, os fatores das constantes normalizadoras sobre os componentes primos do grafo são como implicados em 5.3, mas as constantes normalizadoras para componentes primos não-completos não tem forma fechada. Métodos de Monte Carlo para estimação dessas constantes normalizadoras são discutidos na seção 4.5.

5.4 Verossimilhança para Modelos Não-Decomponíveis

Para modelos não - decomponíveis, as constantes normalizadoras que correspondem aos fatores na equação (5.3) que representam componentes primos não-completos não tem forma fechada. Elas podem ser expressas como integrais sobre os espaços de Σ_{PP} , compatíveis com o conjunto de arestas E do componente primo P. Para simplificar a notação assumimos que P constitui o grafo todo, tal que $\Sigma = \Sigma_{PP}$. Então, temos

$$h(P, \delta, \Phi) = \int_{\Sigma^{E|P}} |\Sigma|^{-\frac{\delta-2}{2}} \exp \left(\frac{-1}{2} \Sigma^{-1} \Phi \right) J(\Sigma^E), \quad (5.8)$$

ou mais simplificadamente, em termos de uma integral sobre $(\Sigma)^{-1} = \Omega$:

$$h(P, \delta, \Phi) = \int_{\Omega^{E|P}} |\Omega|^{\frac{\delta-2}{2}} \exp\left(\frac{-1}{2}\Omega\Phi\right) d\Omega^E. \quad (5.9)$$

Para estimar estas integrais, Dobra *et al.* (2004) usam o método apresentado em Atay-Kayis e Massam (2003), daqui por diante AM05. Eles exploram duas mudanças de variáveis de Ω^E para Φ^E , os elementos livres da matriz triangular superior produzida pela decomposição de Cholesky de Ω ; e de ϕ^E para ψ^E , onde $\psi = \Phi T^{-1}$, e $T'T$ é a decomposição de Cholesky de Φ .

O ponto desta mudança é o que os elementos livres de ψ são normais independentes e raiz quadrada de variáveis aleatórias χ^2 , e são facilmente gerados, os elementos não livres podem ser simplesmente calculados dos elementos livres. A equação 5.9, escrita em termos de ψ , torna-se

$$h(P, \delta, \Phi) = \left(\prod_{i=1}^{|P|} 2^{\frac{\delta+\nu_i}{2}} (2\pi)^{\frac{\nu_i}{2}} \Gamma\left(\frac{\delta+\nu_i}{2}\right) T_{ii}^{\frac{\delta+b_i-1}{2}} \right) E_{\psi^E} (f_T(\psi^E)), \quad (5.10)$$

onde ν_i é o número de vizinhos do nó i subsequente na ordenação de vértices, b_i , é o número total de vizinhos do nó i mais um, e

$$f_T(\psi^E) = \exp\left(\frac{-1}{2} \sum_{(i,j) \notin E, i < j} \psi_{i,j}^2\right) \quad (5.11)$$

O método de AM05 foi construído com algumas das idéias de Roverato (2002) onde o método *importance sampling* para calcular (5.9) é desenvolvido. Roverato usa um método aproximadamente decomponível, com conjunto de arestas E^* que contem E . Para este modelo, a mudança de variáveis de Ω^{E^*} para Φ^{E^*} é realizada, a distribuição de elementos em Φ^{E^*} é facilmente amostrado por causa da decomponibilidade. Se P não é decomponível, Φ^E tem mais elementos forçados que Φ^{E^*} ; entretanto, valores para entradas forçadas compatíveis com E podem ser simplesmente calculadas a partir dos elementos de Φ^{E^*} que correspondem aos elementos livres de Φ^E .

Este método como em Roverato, usa mudança de variáveis, escrevendo a constante normalizadora como uma esperança sobre as variáveis, e usa o *importance sampling* para aproximar a esperança. A amostragem é baseada nas variáveis aleatórias normais multivariadas Z_i . Para um grafo completo com conjunto de arestas E^* , Ω^{E^*} tem uma

densidade de Wishart. Se os parâmetros são δ , Φ , um valor para Ω^{E*} pode ser gerado como $\sum_{i=1}^{\delta} Z_i Z_i'$ onde Z_i tem covariância Φ . Como para ψ e Φ em prévias provas, a restrição em Ω^E traduz em elementos não-livres na coleção de Z_i 's que podem ser computados como funções de elementos livres. Depois modificando os elementos não - livres, as variáveis resultantes podem ser usadas para estimar a esperança desejada.

Um contraste adicional entre os métodos está no tamanho máximo dos componentes primos para um dado número de pontos. Nos métodos de Dellaporta e AM05, o tamanho do componente primo não pode exceder $n - 1$, onde n é o tamanho da amostra. Outro ponto, a submatriz de ψ^E não é necessariamente inversível.

Em todos os exemplos, o método de AM05 é usado para calcular verossimilhanças marginais para grafos não - decomponíveis. Dobra *et al.* (2004) preferem este método ao *importance sampling* porque ele evita as preocupações sobre a eficiência do amostrador *importance*, isto é, até onde a distribuição dos Φ 's gerados dos Z 's podem ser da distribuição desejada quando é baseado na amostragem direta via composição.

5.5 Algoritmo de Monte Carlo via Cadeia de Markov

MCMC é uma das ferramentas mais usadas para exploração de estruturas gráficas. No contexto de modelos grafos Gaussianos, Wong e Carter (2002) usam seus resultados para construir um “amostrador de Gibbs” fixado para grafos decomponíveis, onde cada aresta foi atualizada conforme esta distribuição condicional completa. Estes resultados são também facilmente explorados em um “amostrador Metropolis-Hastings”. Jones *et al.* (2004), constroem três amostradores para comparar o espaço de grafos decomponíveis: Gibbs, Metropolis-Hastings onde a aresta a ser atualizada é escolhida aleatoriamente, e Metropolis-Hastings onde a escolha da adição ou remoção de uma aresta é arbitrária e então uma aresta é selecionada aleatoriamente daquelas apropriadas para cada tipo de mudança. Não houve diferença evidente na performance entre estes algoritmos MCMC relacionados, os resultados apresentados são da adição/remoção do amostrador Metropolis-Hastings.

Jones *et al.* (2004) também implementam a adição/remoção do amostrador Metropolis - Hastings para uma procura não restrita de espaço de grafos. Quando avaliamos uma

proposta envolvendo um grafo não decomponível, o algoritmo descrito na seção 4.5 é usado para avaliar a verossimilhança. A computação é consideravelmente sobrecarregada, ver tabelas 5.1 e 5.2.

Para problemas com até um número moderado de variáveis nos espaços decomponíveis ou espaços não restritos, o espaço a ser explorado é tão grande que uma frequência do grafo na amostra de grafos produzidos não podem ser vistas quando reflete a probabilidade *a posteriori*. A probabilidade *a posteriori* do grafo estimado é baseada na função *a posteriori* normalizadora usando grafos visitados, e estas quantidades refletirão a verdadeira função *a posteriori*. Entretanto, as probabilidades marginais de inclusão de arestas, podem ser vistas como probabilidades *a posteriori*.

5.6 Algoritmo de Busca Estocástica *Shotgun*

Se MCMC é visto meramente como uma ferramenta para visitação de regiões de alta probabilidade do espaço grafo, há certamente outros algoritmos que podem ser utilizados. O seguinte algoritmo é atrativo:

1. Começar com um grafo G .
2. Selecionar vizinhos X_1 aleatórios (grafos diferindo por uma aresta), calcular sua massa *a posteriori* não normalizada, e reter o topo $X_2 \leq X_1$.
3. Dentre os vizinhos de X_2 , propor o i -ésimo Grafo G^i com um novo grafo inicializador com probabilidade *a posteriori* não normalizada p_i^α , onde p_i é a probabilidade *a posteriori* não normalizada do grafo i e α é um parâmetro *annealing*¹.
4. Retornar ao passo 2 e iterar. Manter a lista de todos os melhores grafos X_3 visitados.

Em experimentação com esta abordagem, Jones *et al.* (2004) têm usado $X_1 = X_2 = T$, assim todos os vizinhos são examinados em cada estágio. Este método é referido como Busca Estocástica *Shotgun*, onde em cada passo geramos um número grande de modelos candidatos; “atirando” em todas as direções, e seguindo um ou mais candidatos

¹É um parâmetro que determina o raio de busca no grafo, está relacionado com *Simulated Annealing*

plausíveis. Quando restringimos a grafos decomponíveis, o passo 2 contém um controle para decomponibilidade; grafos não - decomponíveis têm probabilidade *a posteriori* zero.

Diferentemente do algoritmo MCMC, frequências de arestas, pesadas pela probabilidade *a posteriori* do grafo estimado, também podem ser vistas como probabilidades *a posteriori* somente para estender que todas as quantidades *a posteriori* são capturadas no topo dos grafos X_3 .

5.7 Exemplos Simulados

Primeiro consideramos dois exemplos simulados por Jones *et al.* (2004) onde o verdadeiro grafo é conhecido. O primeiro grafo, Figura 5.4, tem 15 nós e é decomponível. O segundo grafo, Figura 5.5, consiste de 12 nós em um único componente primo não-completo. Cada conjunto de dados consiste de 250 observações.

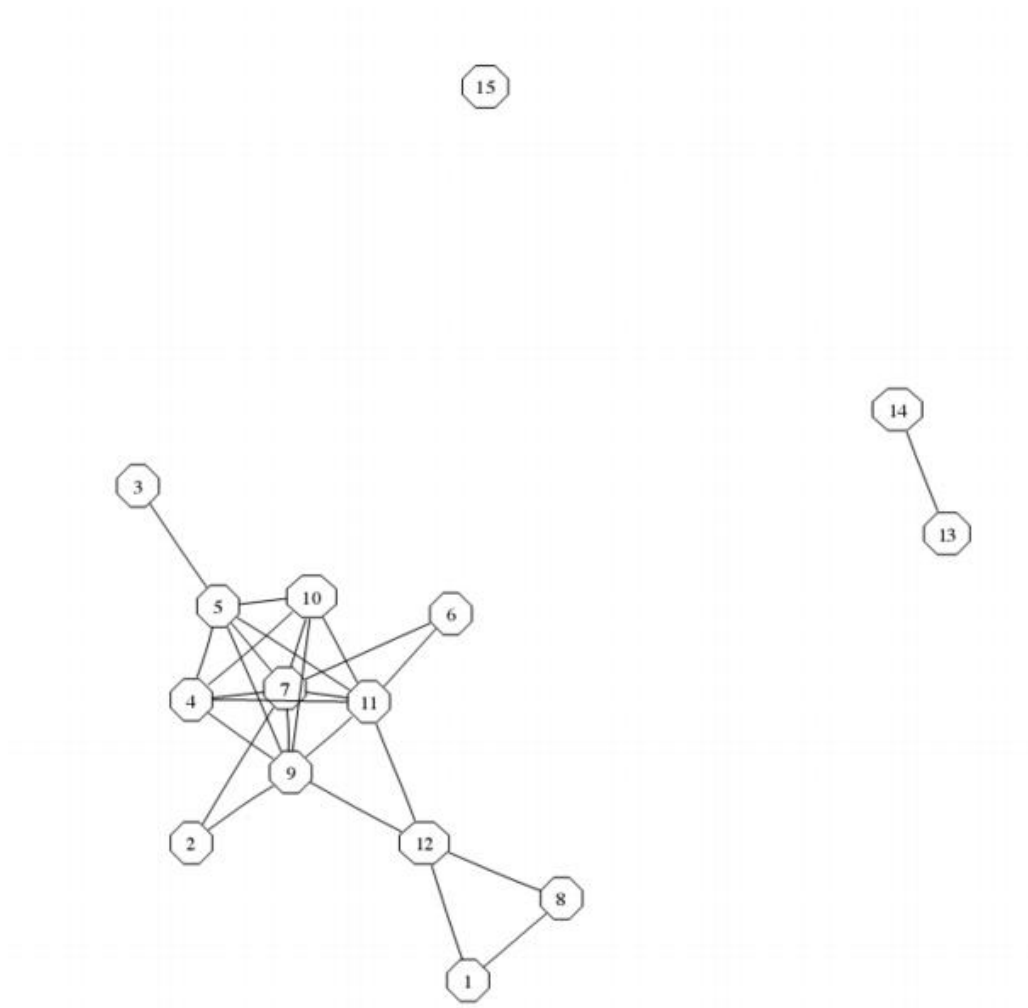


FIGURA 5.4: Grafo decomponível com 15 nós - primeiro exemplo simulado.

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

percorrido de tal forma que tinha o mesmo número de grafos de verossimilhança avaliados pelo Metropolis-Hastings. Cada iteração do algoritmo de Busca Estocástica *Shotgun* inclui avaliação do grafo atual; foram rodadas 10000 iterações de busca estocástica.

Os melhores grafos decomponíveis - são identificados com maior probabilidade *a posteriori* - mostrados na Figura 5.6 e 5.7; os melhores grafos à partir da procura não restrita aparecem nas Figuras 5.8 e 5.9. A comparação de verossimilhança com verdadeiros grafos mostram que cada um desses grafos tem maior suporte de verossimilhança tão bem quanto maior suporte *a posteriori* que o verdadeiro grafo. Também vemos que em geral, as arestas incluídas tem maior probabilidade *a posteriori* estimada do que as não incluídas. Para todos, menos o grafo decomponível com 15 nós, todas as arestas incluídas tem maior probabilidade que as excluídas. (O nó 15 do grafo decomponível inclui uma exceção a isto: a probabilidade mais baixa da aresta incluída tem probabilidade 0.58, enquanto a maior probabilidade da aresta incluída tem probabilidade 0.6). Então, observando a aresta com probabilidade maior em um grafo, isto não resulta em grafos muito diferentes do que, tomando o melhor grafo encontrado.

O grafo mais provável encontrado no caso do grafo de 12 nós, e os casos decomponíveis, foram insensíveis ao ponto de início, o mesmo grafo foi encontrado começando pelo grafo completo. A procura não restrita para o nó 15 começando pelo grafo completo não atingiu a verossimilhança para o topo do grafo mostrado na tabela.

TABELA 5.1: Comparação entre o tempo para rodar os algoritmos, e qualidade do melhor grafo encontrado, para o exemplo de 12 nós. MH-d(nr) refere-se ao Metropolis-Hastings em modelos decomponíveis (não restritos), enquanto o BES-d(nr) refere-se ao método de Busca Estocástica *Shotgun* em modelos decomponíveis (não restritos)

Grafo com 12 nós				
Métodos	Tempo (seg)	Max log <i>posteriori</i>	Nº Grafos visitados antes de encontrar o melhor	Tempo p/ encontrar o melhor
MH-d	36	-2591,18	912	1
BES-d	183	-2591,18	792	2
MH-nr	15.220	-2590,94	415	2
BES-nr	2273	-2590,94	13,266	5

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

TABELA 5.2: Comparação entre o tempo para rodar os algoritmos, e qualidade do melhor grafo encontrado, para o exemplo de 15 nós. MH-d(nr) refere-se ao Metropolis-Hastings em modelos decomponíveis (não restritos), enquanto o BES-d(nr) refere-se ao método de Busca Estocástica *Shotgun* em modelos decomponíveis (não restritos)

Grafo com 15 nós				
Métodos	Tempo (seg)	Max log <i>posteriori</i>	Nº Grafos visitados antes de encontrar o melhor	Tempo p/ encontrar o melhor
MH-d	93	15633,76	349,484	36
BES-d	234	15633,76	33,495	9
MH-nr	513.077	15633,83	66,425	309,222
BES-nr	5930	15636,38	82,845	112

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

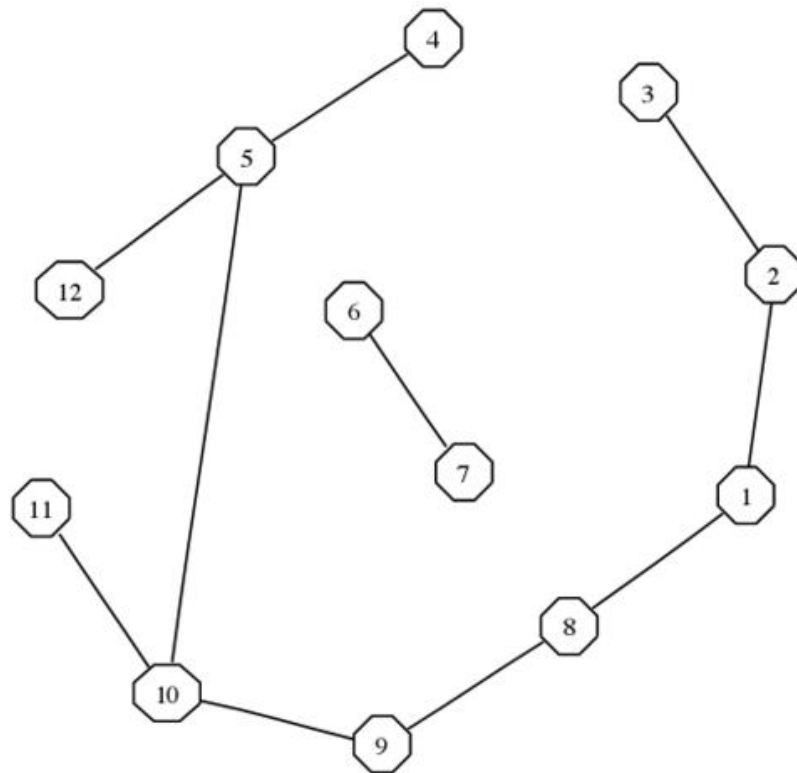


FIGURA 5.6: Grafo de maior log *posteriori* para o exemplo com 12 nós quando a busca é restrita para modelos decomponíveis.

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

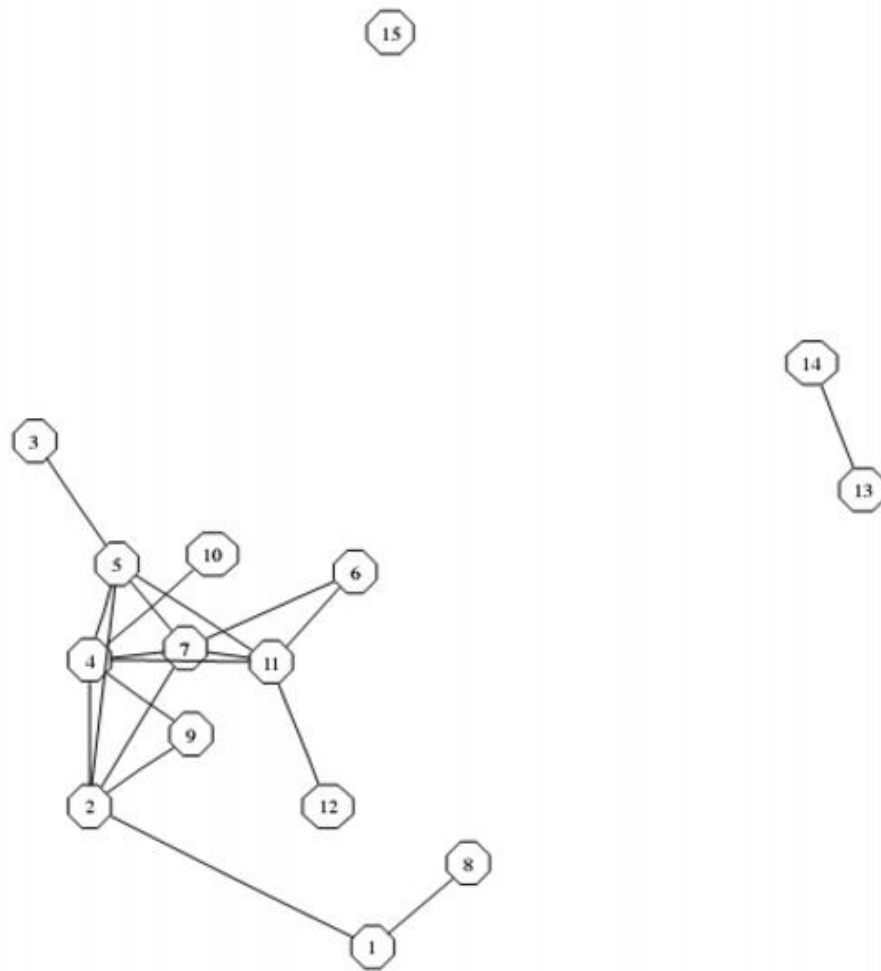


FIGURA 5.7: Grafo de maior log *posteriori* para o exemplo com 15 nós quando a busca é restrita para modelos decomponíveis.

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

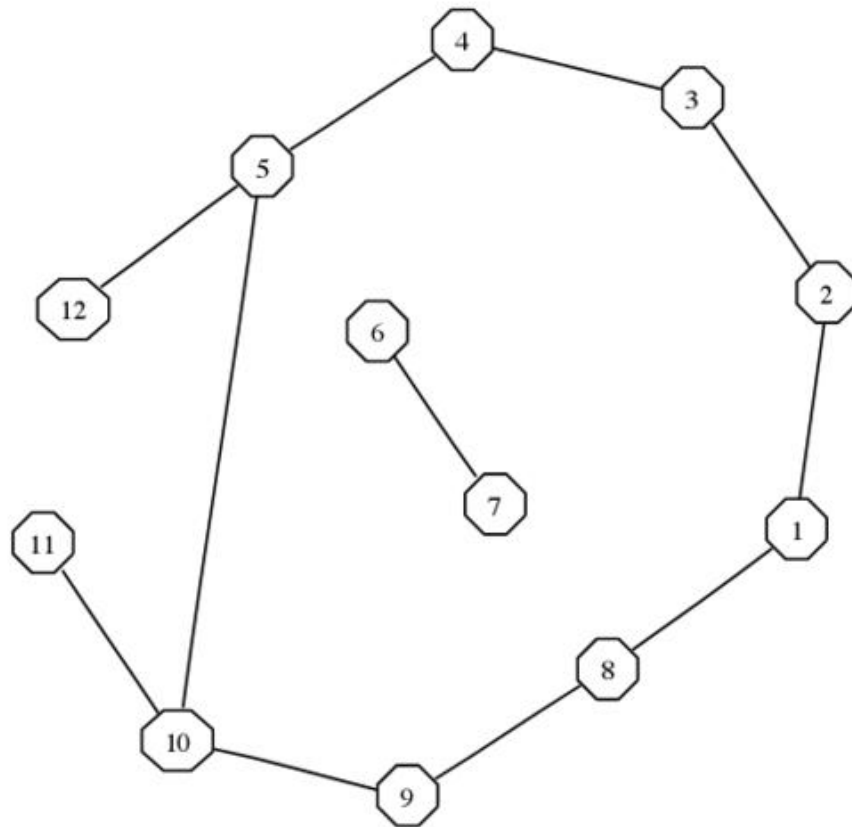


FIGURA 5.8: Grafo de maior log *posteriori* para o exemplo com 12 nós quando a busca é não restrita.

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

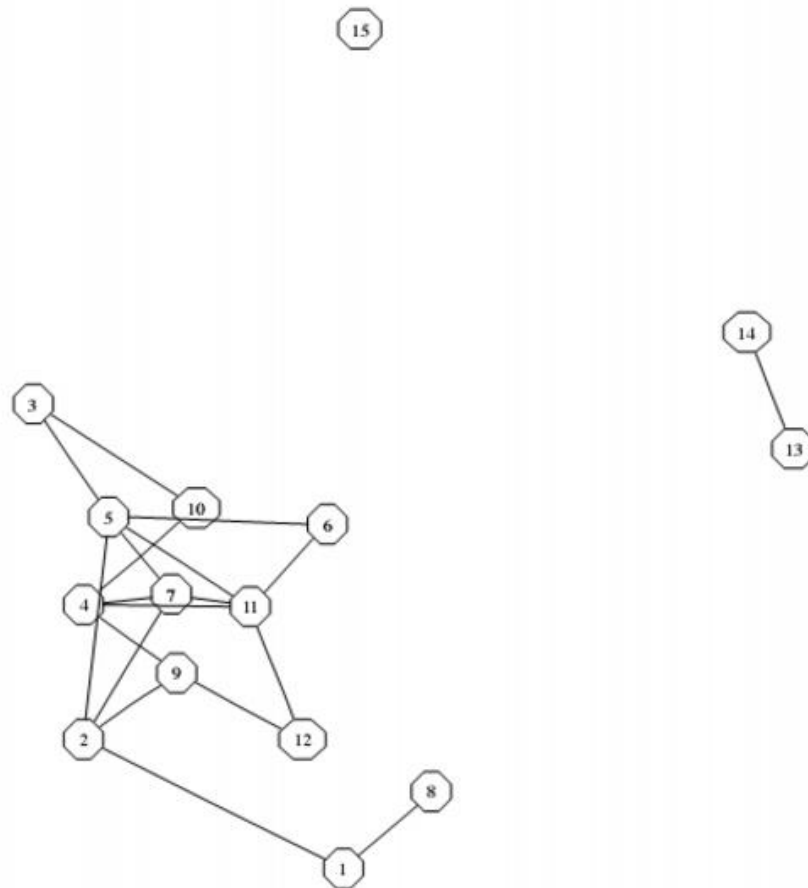


FIGURA 5.9: Grafo de maior log *posteriori* para o exemplo com 15 nós quando a busca é não restrita.

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

5.8 Exemplo: Dados de Expressão Gênica

Um problema desafiante é a análise de dados de expressão gênica, para $p = 150$ genes associados com o estrógeno receptor, tomado de $n = 49$ indivíduos; os dados vem do estudo de West em (2001). Os dados foram padronizados e a *priori* especificada com $\delta = 3$, $\tau = 4$. Neste contexto, a *priori* pode ser interpretada de tal forma que em média cada gene tenha interações importantes com um número relativamente pequeno de outros genes. Neste exemplo, Jones *et al.* (2004), adicionam à *priori* do modelo decomponível a restrição do tamanho do clique não exceder a $n - 1$, para manter a identificabilidade do modelo. Para o modelo não - restrito, conta que cada componente primo tem menos $n - 1$ vértices.

Os resultados dos três algoritmos são mostrados na Tabela 5.3. O tempo é dado em hora. A procura não restrita do Metropolis-Hastings mostrou performance inferior, por isso não foi usada. Os melhores resultados para o algoritmo de Busca Estocástica *Shotgun* foram obtidos quando um parâmetro *annealing* de 50 foi usado. Neste exemplo vemos que, até no caso decomponível o algoritmo de Busca Estocástica *Shotgun* encontra os melhores grafos.

Um parâmetro grande foi também usado para o algoritmo de Busca Estocástica *Shotgun* para o caso não restrito. No entanto, neste caso o parâmetro não elimina a estocasticidade da procura.

Começando de um grafo vazio até a estimada *log posteriori*, o melhor grafo encontrado tem *log posteriori* -9364,67, pior que o melhor grafo decomponível. A tabela mostra o melhor grafo encontrado começando pelo melhor grafo decomponível (a estimativa final da *log posteriori* para este grafo foi rodado com iterações suficientes, coloca o desvio padrão abaixo de 1 unidade da *log verossimilhança*). Um total de 10 ciclos da avaliação de todos os vizinhos foi feito. Como estes grafos foram grafos decomponíveis “fechados”, o tempo requerido para avaliá-los foi também reduzido *versus* grafos com números similares de arestas produzidas pela procura começando pelo grafo vazio.

TABELA 5.3: Comparação entre o tempo para rodar os algoritmos, e qualidade do melhor grafo encontrado, para o exemplo de expressão gênica. * Começando do melhor grafo decomponível encontrado. MH-d refere-se ao Metropolis-Hastings em modelos decomponíveis, enquanto o BES-d(nr) refere-se ao método de Busca Estocástica *Shotgun* em modelos decomponíveis (não restritos)

Exemplo Expressão Gênica				
Métodos	Tempo (h)	Max log <i>posteriori</i>	Nº Grafos visitados antes de encontrar o melhor	Tempo p/ encontrar o melhor
MH-d	18,02	-9417,97	100466,818	6,51
BES-d	0,03	-9260,84	1698,600	0,03
BES-nr	6,29*	-9227,68	44700	3,39

Fonte: Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004)

5.9 Discussão

A exploração de modelos para encontrar grafos de alta probabilidade *a posteriori* pode ser sucessivamente realizada usando procura direta, tal como com o método de Busca Estocástica *Shotgun*. O tradicional MCMC é competitivo para grafos relativamente pequenos.

Grandes componentes primos induzem uma responsabilidade computacional grande, via estimação pelo método Monte Carlo das constantes normalizadoras, estimação que pode ser muito instável quando a dimensão aumenta. Outros métodos são necessários para esse problema computacional. A busca local de grafos não restritos, em torno de “bons” grafos decomponíveis ou outros grafos candidatos, é possível para 150 variáveis e representam uma estratégia promissora.

No caso de buscas não restritas, novas abordagens teóricas e métodos são necessários para mostrar a capacidade de estimar a constante normalizadora, associada com componentes primos não completos em uma representação de árvore de junções. Flores, Gomes e Olesen (2003) endereçam este problema ao contexto de grafos orientados, estes resultados poderiam ser aplicados para provar a caracterização de trocas análogas de componentes primos, para cliques nos grafos decomponíveis usados em Giudici e Green (1999).

Dobra, Hans, Jones, Nevins, Yao e West (2004) têm recentemente introduzido como abordagem a construção de modelos grafos Gaussianos de alta - dimensão, construindo a nova distribuição conjunta triangular de regressões representando as relações entre as variáveis. Esta abordagem é relacionada à rede de dependência de Heckerman, Chickering, Meek, Rounthwaite e Kadie (2000).

Enfoques construtivos baseados em regressões requerem uma ordenação de variáveis, e isto pode evidentemente ter um impacto maior na eficiência computacional e nas regiões do modelo grafo visitado. No entanto, os experimentos com métodos MCMC e busca estocástica relacionados aos métodos MCMC levá-nos a concluir que é necessário um controle da escala, além de dimensões moderadas. O exemplo em Dobra (2004) consiste de dados de expressão gênica sob 12000 genes, indicando que o enfoque está pelo menos implementável com conjuntos muito grandes de variáveis; e o exemplo aparentemente identifica grafos que são, no contexto biológico, interpretáveis.

Um conceito geral que vê muitos progressos está no desenvolvimento de métodos capazes de rotineiramente gerar “grandes” saltos no modelo grafo e, então, integrar com MCMC ou busca baseada em métodos locais móveis.

Esses experimentos tem usado pesadamente o *cluster Beowulf*, com o aumento ao acesso de grandes *clusters* para computação, a estatística computacional tenta encontrar a habilidade para explorar modelos de alta-dimensão baseada na tecnologia que integra-se à pesquisa diária.

5.10 Simulação

Foi feito um estudo de simulação utilizando o metropolis-hastings para grafos decomponíveis baseados em Jones *et al.* (2004). Primeiro, consideramos o exemplo das notas matemáticas, visto na seção 2.4. Os dados consistem de 5 variáveis e 88 observações, com parâmetros $\tau = 0.0004$, $\delta = 3.0$ e 1000 iterações.

O melhor grafo decomponível foi identificado com maior *log posteriori* de -1920.95, e foi como esperado (conferir Figura 3.4).

Depois analisamos dados de expressão gênica para $p = 5963$ genes associados ao câncer de mama, tomado de 32 indivíduos; os dados vêm do estudo de Veer *et al.* (2002). Os parâmetros utilizados foram $\tau = 0.0004$, $\delta = 3.0$ com 1000000 iterações. O melhor grafo decomponível foi identificado com a maior *log posteriori* de 47391.6. Podemos observá-lo na Figura 5.10

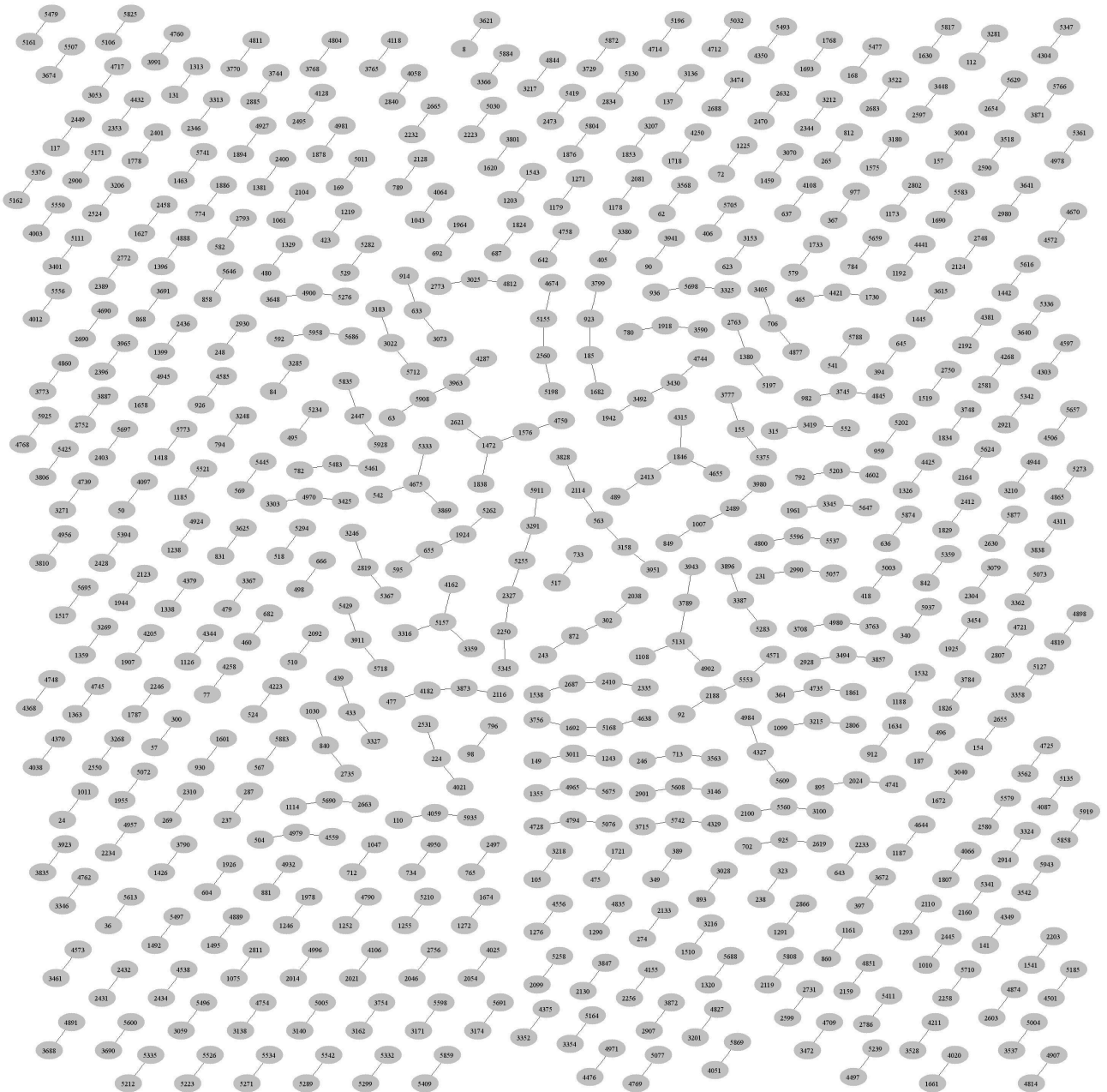


FIGURA 5.10: Melhor grafo decomponível encontrado pelo metropolis-hastings.

Retiramos as arestas duplas, devido a simplicidade delas, e observamos então o grafo resultante através da Figura 5.11

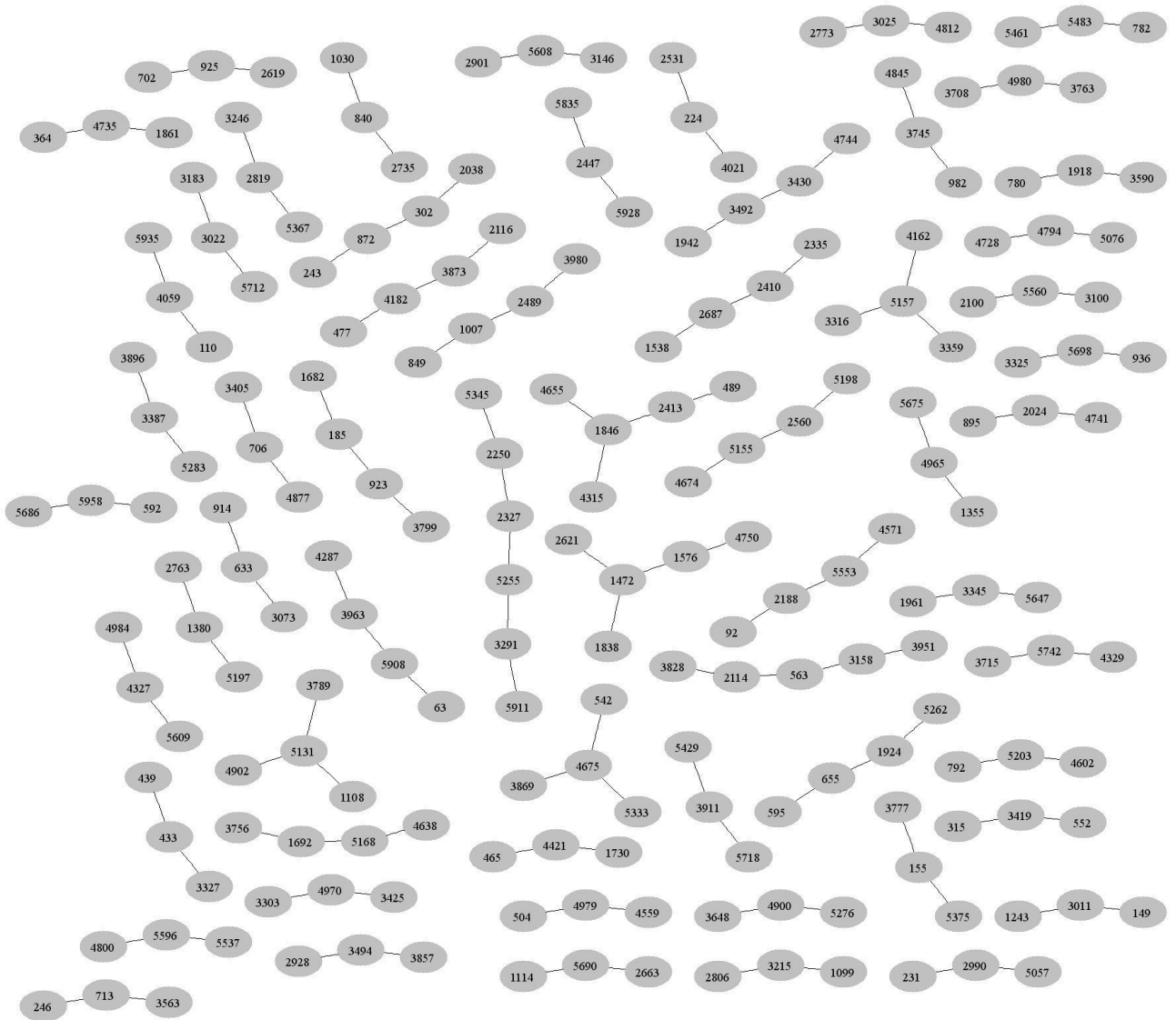


FIGURA 5.11: Melhor grafo decomponível encontrado pelo metropolis-hastings, sem as arestas duplas.

Observamos que o Metropolis-Hastings é um algoritmo eficiente no trabalho com dados de dimensões grandes, trazendo resultados satisfatórios.

Porém, não foi possível comparar os resultados obtidos do Metropolis com o algoritmo de Busca Estocástica *Shotgun*, pois este necessita de um investimento computacional bem maior, que não tínhamos disponível.

Capítulo 6

Considerações Finais

Nesta dissertação apresentamos uma introdução aos Modelos Grafos Probabilísticos, que são utilizados para visualizar a estrutura de dependência e independência condicional existente entre as variáveis de um vetor aleatório, através de arestas e vértices.

Uma introdução à Genética é abordada no Capítulo 1, dado que este trabalho é voltado para variáveis usadas em expressão gênica. Neste capítulo descrevemos a formação do DNA, além do segmento de DNA que contém informações para produção de proteínas, ou seja, os genes. Os genes, para muitos cientistas, trazem a explicação para inúmeros problemas genéticos.

No Capítulo 2, fizemos uma revisão da teoria dos grafos. Descrevemos os Modelos Grafos Gaussianos e apresentamos, baseados em Whittaker (1990), um exemplo de aplicação da metodologia a um conjunto de dados, mostrando a relação entre as notas de alunos em cinco disciplinas. Com base em Edwards (1995), apresentamos modelos grafos para tabelas de contingência para multi-entradas, baseadas em amostragem da distribuição Multinomial, e um exemplo onde foi possível identificar o paradoxo de Simpson.

Motivados pelo trabalho de Dobra, Hans, Jones, Nevins e West (2003), no Capítulo 3, discutimos a estrutura teórica e metodologia para modelos de grande escala, visando a exploração de associações entre genes utilizando dados de expressão gênica.

Baseados no método de Dobra *et al.* (2003), para modelar a estrutura de covariância de distribuições de alta dimensão, com foco na estrutura esparsa e no trabalho de Seokho Lee (2006), desenvolvemos um programa que visa captar relações entre variáveis.

Aplicamos ao programa dois conjuntos de dados, o primeiro referente as cinco notas matemáticas relacionadas a 88 estudantes, presente na seção 2.4. Em relação a este conjunto, já sabíamos o modelo grafo resultante, e pudemos confirmá-lo (ver, por exemplo, Figuras 3.3 e 3.4).

Para o segundo conjunto de dados, simulamos uma matriz de valores normais, com 10 colunas e 15 linhas, e colocamos uma associação entre as variáveis 3 e 4, 5 e 6, 7 e 8, 9 e 10. Observamos que as associações apareceram, desta forma verificamos que o algoritmo tem um bom funcionamento, Figuras 3.5 e 3.6).

No Capítulo 5, tendo como base Fronk (2002), foi proposta uma abordagem bayesiana usando *Reversible Jump* MCMC (RJMCMC), introduzido por Green (1995), este método é capaz de controlar a variação de dimensão do espaço de busca, e isto é interessante porque a seleção de modelos grafos enfrenta o problema do espaço de busca aumentar exponencialmente com o número de variáveis incorporadas na análise. Devido ao enorme número de possíveis modelos, não é viável julgá-los por um critério de qualidade de ajuste como o AIC ou o BIC e encontrar o melhor modelo com respeito a estes critérios. Este capítulo nos ajudou a ampliar nosso conhecimento, verificando mais uma metodologia que pode ser aplicada para casos em que existam grandes números de variáveis envolvidas no modelo.

Descrevemos no Capítulo 4 os modelos grafos decomponíveis, visando a eficiência computacional que surge através da decomposição dos espaços amostrais a partir de subconjuntos de variáveis (vértices nos grafos) baseada nas suas relações gráficas. Artigos recentes envolvem questões de melhoramentos dos métodos computacionais para estimação de modelos grafos Gaussianos não-orientados. Giudici e Green (1999) mostram condições computáveis fáceis e eficientes para determinar se ao adicionarmos uma aresta a um grafo, mantemos a propriedade de decomposição.

Foi feito um estudo de simulação utilizando o metropolis-hastings para grafos decomponíveis baseados em Jones *et al.* (2004). Consideramos, primeiramente o exemplo das notas matemáticas, visto na seção 2.4, com parâmetros $\tau = 0.0004$, $\delta = 3.0$ e 1000 iterações.

O melhor grafo decomponível foi identificado com maior log *posteriori* de -1920.95, e foi como esperado (conferir Figura 3.4).

Depois analisamos dados de expressão gênica para $p = 5963$ genes associados ao câncer de mama, tomado de 32 indivíduos; os dados vêm do estudo de Veer *et al.* (2002). Os parâmetros utilizados foram $\tau = 0.0004$, $\delta = 3.0$ com 1000000 iterações. O melhor grafo decomponível foi identificado com a maior log *posteriori* de 47391.6 (verificar Figura 5.10). Retiramos as arestas duplas, devido a simplicidade delas, e observamos então o grafo resultante através da Figura 5.11.

Observamos que o Metropolis-Hastings é um algoritmo eficiente no trabalho com dados de dimensões grandes, trazendo resultados satisfatórios.

Pudemos verificar que apesar do tema abordado nesta dissertação ser de grande importância, visto que existem muitos problemas em que o conjunto de dados compreende um grande número de variáveis, ainda faltam referências consistentes a este respeito, principalmente quando o intuito é trabalhar com modelos grafos, mas estas metodologias apontadas e verificadas neste trabalho têm sido satisfatórias.

Enfim, esta dissertação aborda três métodos de seleção de variáveis para modelos grafos, baseados em modelos esparsos, *Reversible Jump* e modelos decomponíveis. Cada um com suas características, como discutidas no texto.

Apêndice A

Definições

[1] *Critério da Fatoração*: As variáveis aleatórias Y e Z são condicionalmente independentes dado X , $Y \perp\!\!\!\perp Z | X$, se, e somente se, existem funções g e h tal que

$$f_{XYZ}(x, y, z) = g(x, y)h(x, z) \text{ para todo } y \text{ e } z \text{ e todo } x \text{ com } f_X(x) > 0.$$

[2] *Informação Divergência* (A divergência entre duas distribuições Normais): Se $f(x) = n(x; \mu_1, V_1)$ e $g(x) = n(x; \mu_2, V_2)$ então

$$I(f; g) = \frac{1}{2}[\mu_1 - \mu_2, V_2^{-1}(\mu_1 - \mu_2)] + \frac{1}{2}tr(V_1V_2^{-1}) - \frac{1}{2}logdet(V_1V_2^{-1}) - \frac{k}{2}.$$

[3] *Moralização*: O grafo moral de um GAO, é o grafo não orientado obtido do GAO, pela adição de arestas entre os pares de vértices não adjacentes que tem um sucessor em comum (vértices v e w têm um sucessor em comum se existe um vértice x com (v, x) no conjunto de arestas e (w, x) no conjunto de arestas), e então trocam-se as arestas direcionadas pelas não orientadas. Caso não exista variáveis que têm em comum o mesmo sucessor, trocam-se simplesmente as arestas direcionadas pelas não orientadas. Por exemplo, se considerarmos o GAO

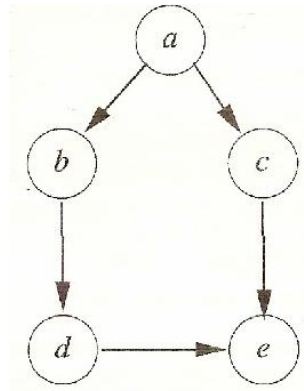


FIGURA A.1: Grafo Acíclico Orientado.

O grafo moral correspondente a A.1 é

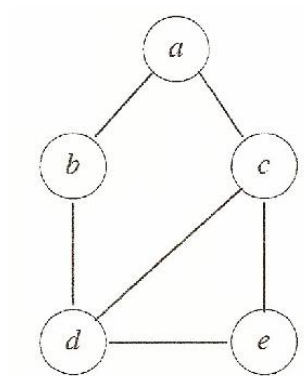


FIGURA A.2: Grafo moral correspondente a A.1.

Apêndice B

Linhas de Comandos

Programa para modelos grafos esparsos - conjunto de dados referente as notas matemáticas e matriz de valores normais

```
a=randn(150,15);
a(:,8)=a(:,7)+0.1*rand(150,1);
a(:,10)=a(:,9)+0.1*rand(150,1);
a(:,6)=a(:,5)+0.1*rand(150,1);
a(:,4)=a(:,3)+0.1*rand(150,1);
dados=a;
tau=1; # define tau
delta=0.01; #define delta
n=size(dados,1); #numero de linhas do dados
p=size(dados,2); #numero de colunas do dados
beta=1/(p-1); #calcula beta
P=sparse(p,p); #cria matriz sparsa dos possíveis preditores
for i=1:p #faz regressão e guarda preditores em P
    vet = [1 : i - 1, i + 1 : p];
    [b, se, pval, inmodel] = stepwisefit(dados(:,vet),dados(:,i),'display','off');
    P(i,vet)=inmodel;
end
Mgra=sparse(p,p); #matrix que guarda os preditores da variáveis selecionadas
P3=P; #cria outra matrix com preditores para não perder o original
```

```

Avet=1:p; #cria vetor com variaveis que serão utilizadas para fazer novas regressões
posteriormente
MSC=[];ind=[];MML=[];SC=[]; #cria variaveis para armazenamento temporario
for contador= 1:p #Cria vetor O (VV) com as variaveis ordenadas
for i=1:p #calcula as verossimilhanças e os scores para cada xi e seleciona o maior score
P2=P3;
P2(:,i)=zeros(size(P2,1),1); #remove os xi dos possíveis preditores para calcular as verossim-
ilhanças
for count=1:numel(Avet)
j=Avet(count); #calcula apenas para as variáveis que tem xi como preditora
y=dados(:,j);
aux=find(P2(j,)==1);
k=numel(aux);
if k==0
M=1;
q=y'*y;
else
z=dados(:,aux);
M=sparse(1:k,1:k,tau)+(z'*z);
q=y'*y-y'*z*inv(M)*z'*y;
end
dM=det(M);
numerador=tauk/2 * gamma((n + delta + k)/2);
denominador=gamma((delta + k)/2) * sqrt(dM) * (1 + q/tau)((n+delta+k)/2);
MML(j)=numerador/denominador * (beta/(1 - beta))(k-1) * 1030;
end
MML2=MML;
if contador =1
MML2(VV)=1;
end
SC(i)=prod(MML2);
end

```

```

if contador =1
SC(VV)=0;
end
MSC=find(SC==max(SC),1,'first');
Mgra(contador,:)=P3(MSC,:);
VV(contador)=MSC;
Avet=find(P3(:,MSC)==1);
P3(:,MSC)=sparse(zeros(size(P3,1),1));
end

```

Programa para modelos grafos decomponíveis - metropolis-hastings

```

# include <stdio.h>
# include <stdlib.h>
# include <string.h>
# include <limits.h>
# include "sparsegraph.h"
# include "graphlist.h"
# define max(a,b) ((a > b) ? a:b)
extern void CheckPointer(void* pointer);
extern int choose(int m, int n);
# checks whether v is an element of the set aset.
# The elements of aset are sorted in ascending order.
# aset has n elements in it.
int isElement(int v,short* aset,int n)
{
int i;
for(i=0;i < n;i++)
{
if(v==aset[i])
{
return 1;
}
}
}

```

```
if(v < aset[i])
{
break;
}
}
return 0;
}
# for qsort
int numeric(const void *p1, const void *p2)
{
return(*((short*)p1) - *((short*)p2));
}
#class Graph::Begins
Graph::Graph(
{
maxNumberOfNeighbors = 0;
maxCliqueDimens = 0;
nVertices = 0;
Cliques = NULL;
CliquesDimens = NULL;
nCliques = 0;
Adjacency = NULL;
AdjacencyDimens = NULL;
cliqueTreeAdjacency = NULL;
cliqueTreeAdjacencyDimens = NULL;
ordering = NULL;
Separators = NULL;
SeparatorsDimens = NULL;
nSeparators = 0;
whichSeparator = NULL;
return;
}
```

```
#the destructor of the graph does
#not free any memory; make sure you
#free the memory with ResetGraph()
#before the destructor is called Graph:: Graph()
{
maxNumberOfNeighbors = 0;
maxCliqueDimens = 0;
nVertices = 0;
nCliques = 0;
nSeparators = 0;
return;
}
# This function does NOT allocate memory
# for the cliques, separators or the clique tree!!
#To save memory, we allocate it
void Graph::InitGraph(int n,int maxneighbors,int maxcliquedimens)
{
int i, j;
maxNumberOfNeighbors = maxneighbors+1;
maxCliqueDimens = maxcliquedimens;
nVertices = n;
#alloc the matrix of vertices
Adjacency = new short*[nVertices];
CheckPointer(Adjacency);
memset(Adjacency,0,nVertices*sizeof(short*));
for(i=0; i < n; i++)
{
Adjacency[i] = new short[maxNumberOfNeighbors];
CheckPointer(Adjacency[i]);
memset(Adjacency[i],0,maxNumberOfNeighbors*sizeof(short));
}
AdjacencyDimens = new short[nVertices];
```



```

memset(AdjacencyDimens,0,nVertices*sizeof(short));
ordering = new short[nVertices];
CheckPointer(ordering);
memset(ordering,0,nVertices*sizeof(short));
return;
}
void Graph::WriteInfo(FILE* out)
{
int i, j;
for(i=0; i < nCliques; i++)
{
fprintf(out, "Clique for(j=0; j < CliquesDimens[i]; j++)
{
fprintf(out, "}
fprintf(out, "");
}
int count = 0;
for(i=0;i < nCliques;i++)
{
for(j=0;j < cliqueTreeAdjacencyDimens[i];j++)

if(i < cliqueTreeAdjacency[i][j])
{
count++;
fprintf(out,"Edge count,
i+1,cliqueTreeAdjacency[i][j]+1);
}
}
}
for(i=0; i < nSeparators; i++)
{
fprintf(out, "Separator for(j=0; j < SeparatorsDimens[i]; j++)

```

```
{
fprintf(out, "}
fprintf(out, "");
}
return;
}
void Graph::PrintAdjacency(FILE* out)
{
int i,j;
for(i=0;i < nVertices;i++)
{
fprintf(out,"for(j=0;j < AdjacencyDimens[i];j++)
{
fprintf(out,"}
fprintf(out,"");
}
return;
}
void Graph::GenerateCliques()
{
int i, j, k, p, r;
int n = nVertices;
short* clique = new short[nVertices];
CheckPointer(clique);
short* LRound = new short[nVertices];
CheckPointer(LRound);
#allocate a graph to store the cliques
LPList mycliques = ::InitGraph(nVertices);
#alocate a graph to store the clique tree
LPList mytree = ::InitGraph(nVertices);
#clean memory
memset(clique,0,nVertices*sizeof(short));
```

```
memset(LRound,0,nVertices*sizeof(short));
int v, vk;
int PrevCard = 0;
int NewCard;
int s = -1;
for(i=n-1;i >= 0;i-)
{
NewCard = -1;
#choose a vertex v...
for(j=0;j < n;j++)
{
#test vertex j
if(LRound[j]==0)
{
int maxj = 0;
for(r=0;r < AdjacencyDimens[j];r++)
{
if(LRound[Adjacency[j][r]])
{
maxj++;
}
}
if(maxj > NewCard)
{
v = j;
NewCard = maxj;
}
}
}
if(NewCard == -1) break;
ordering[v] = i;
if(NewCard <= PrevCard)
```

```

{
#begin new clique
s++;
for(r=0;r < AdjacencyDimens[v];r++)
{
if(LRound[Adjacency[v][r]])
{
::AddElement(mycliques+s,Adjacency[v][r]);
#Cliques[s][CliquesDimens[s]] = Adjacency[v][r];
#CliquesDimens[s]++;
}
}
if(NewCard != 0)
{
LPList pady = mycliques[s].Next;
#at least one vertex should have
#been added to this clique
vk = pady->vertex;
k = ordering[vk];
pady = pady->Next;
while(NULL!=pady)
{
if(ordering[pady->vertex]< k)
{
vk = pady->vertex;
k = ordering[vk];
}
pady = pady->Next;
}
#get edge to parent
p = clique[vk];
#add edge (s,p) to the clique tree

```

```

::AddEdge(mytree,s,p);
}
}
clique[v] = s;
::AddElement(mycliques+s,v);
#Cliques[s][CliquesDimens[s]] = v;
#CliquesDimens[s]++;
LRound[v] = 1;
PrevCard = NewCard;
}
nCliques = s+1; delete[] clique; clique = NULL;
delete[] LRound; LRound = NULL;
#generate the vectors representing the cliques
#first, allocate the memory
Cliques = new short*[nCliques];
CheckPointer(Cliques);
memset(Cliques,0,nCliques*sizeof(short*));
CliquesDimens = new short[nCliques];
CheckPointer(CliquesDimens);
memset(CliquesDimens,0,nCliques*sizeof(short));
#transform the graph in vectors
for(i=0;i < nCliques;i++)
{
LPList pady = mycliques[i].Next;
while(NULL!=pady)
{
CliquesDimens[i]++;
pady = pady->Next;
}
Cliques[i] = new short[CliquesDimens[i]];
pady = mycliques[i].Next;
j=0;

```

```

while(NULL!=pady)
{
Cliques[i][j] = pady->vertex;
j++;
pady = pady->Next;
}
}
#generate the vectors representing the cliques tree
#first allocate the memory
cliqueTreeAdjacency = new short*[nCliques];
CheckPointer(cliqueTreeAdjacency);
memset(cliqueTreeAdjacency,0,nCliques*sizeof(short*));
cliqueTreeAdjacencyDimens = new short[nCliques];
memset(cliqueTreeAdjacencyDimens,0,nCliques*sizeof(short));
#as before, transform the graph in vectors
for(i=0;i < nCliques;i++)
{
LPList pady = mytree[i].Next;
while(NULL!=pady)
{
cliqueTreeAdjacencyDimens[i]++;
pady = pady->Next;
}
cliqueTreeAdjacency[i] = new short[cliqueTreeAdjacencyDimens[i]];
pady = mytree[i].Next;
j=0;
while(NULL!=pady)
{
cliqueTreeAdjacency[i][j] = pady->vertex;
j++;
pady = pady->Next;
}
}

```

```

}
#alloc the matrix that tells us
whichSeparator = new short*[nCliques];
for(i=0;i < nCliques;i++)
{
whichSeparator[i] = new short[nCliques];
memset(whichSeparator[i],0,nCliques*sizeof(short));
}
#PARTIALLY allocate memory for the separators
nSeparators = nCliques-1;
Separators = new short*[nSeparators];
CheckPointer(Separators);
memset(Separators,0,nSeparators*sizeof(short*));
SeparatorsDimens = new short[nSeparators];
CheckPointer(SeparatorsDimens);
memset(SeparatorsDimens,0,nSeparators*sizeof(short));
#generate the corresponding separators
GenerateSeparators();
#clean memory
DeleteGraph(mycliques,nVertices);
delete[ ] mycliques; mycliques = NULL;
DeleteGraph(mytree,nVertices);
delete[ ] mytree; mytree = NULL;
return;
}
void Graph::GenerateSeparators()
{
int i1,i2;
int j, k;
int FirstClique, SecondClique;
int v;
nSeparators = 0;

```

```

for(i1=0;i1 < nCliques;i1++)
{
for(i2=0;i2 < cliqueTreeAdjacencyDimens[i1];i2++)
{
if(i1 < cliqueTreeAdjacency[i1][i2])
{
FirstClique = i1;
SecondClique = cliqueTreeAdjacency[i1][i2];
LPList mysep = new List;
mysep->vertex = 0;
mysep->Next = NULL;
for(j=0;j < CliquesDimens[FirstClique];j++)
{
v = Cliques[FirstClique][j];
for(k=0;k < CliquesDimens[SecondClique];k++)
{
if(v==Cliques[SecondClique][k])
{
AddElement(mysep,v);
#Separators[nSeparators][SeparatorsDimens[nSeparators]] = v;
SeparatorsDimens[nSeparators]++;
break;
}
}
if(v < Cliques[SecondClique][k])
{
break;
}
}
}
if(0==SeparatorsDimens[nSeparators])
{
printf("NULL SEPARATOR");
}
}
}
}

```



```

exit(1);
}
Separators[nSeparators] = new short[SeparatorsDimens[nSeparators]];
#transform the list in a vector
LPList pady = mysep->Next;
j=0;
while(NULL!=pady)
{
Separators[nSeparators][j] = pady->vertex;
j++;
pady = pady->Next;
}
DeleteList(mysep);
delete mysep; mysep = NULL;
#qsort((void*)Separators[nSeparators],SeparatorsDimens[nSeparators],sizeof(short),numeric);
whichSeparator[FirstClique][SecondClique] = nSeparators;
whichSeparator[SecondClique][FirstClique] = nSeparators;
nSeparators++;
}
}
}
if(nSeparators!=nCliques-1)
{
printf("INCORRECT NUMBER OF SEPARATORS nSeparators,nCliques-1);
exit(1);
}
return;
}
void Graph::ResetGraph()
{
int i;
for(i=0;i < nCliques;i++)

```

```

{
delete[ ] whichSeparator[i];
whichSeparator[i] = NULL;
}
delete[ ] whichSeparator;
whichSeparator = NULL;
for(i=0;i < nCliques;i++)
{
delete[ ] cliqueTreeAdjacency[i];
cliqueTreeAdjacency[i] = NULL;
}
delete[ ] cliqueTreeAdjacency;
cliqueTreeAdjacency = NULL;
delete[ ] cliqueTreeAdjacencyDimens;
cliqueTreeAdjacencyDimens = NULL;
for(i=0;i < nCliques;i++)
{
delete[ ] Cliques[i];
Cliques[i] = NULL;
}
delete[ ] Cliques; Cliques = NULL;
delete[ ] CliquesDimens; CliquesDimens = NULL;
delete[ ] ordering; ordering = NULL;
for(i=0;i < nSeparators; i++)
{
delete[ ] Separators[i];
Separators[i] = NULL;
}
delete[ ] Separators; Separators = NULL;
delete[ ] SeparatorsDimens; SeparatorsDimens = NULL;
#ifdef DESTRUCTOR
printf("destructor called :: BEGINS");
printf("nCliques = "); printf("nVertices = "); for(i=0;i < nVertices;i++)

```

```
{
#printf("i = delete[ ] Adjacency[i];
Adjacency[i] = NULL;
}
delete[ ] Adjacency; Adjacency = NULL;
#printf("got here");
delete[ ] AdjacencyDimens; AdjacencyDimens = NULL;
#printf("but not here");
#printf("destructor called :: ENDS");
nVertices = 0;
return;
}
void Graph::AddEdge(int vi,int vj)
{
Adjacency[vi][AdjacencyDimens[vi]] = vj;
AdjacencyDimens[vi] += 1;
Adjacency[vj][AdjacencyDimens[vj]] = vi;
AdjacencyDimens[vj] += 1;
return;
}
void Graph::SortAdjacency()
{
int i;
for(i=0;i < nVertices;i++)
{
if(0==AdjacencyDimens[i]) continue;
qsort((void*)Adjacency[i],AdjacencyDimens[i],sizeof(short),numeric);
}
return;
}
#class Graph::Ends
```

Referências Bibliográficas

- [1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.
- [2] Albert, J. H. e Chib, S. (1993). Bayesian Analysis of Binary Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669-679.
- [3] Andersson, S. A., Madigan, D., Perlman, M. D. e Richardson, T. (1997). A Characterization of Markov Equivalence Classes for Acyclic Digraphs. *The Annals of Statistics*, **25**, 505-541.
- [4] Andersson, S. A., Madigan, D., Perlman, M. D. e Richardson, T. (1999). Graphical Markov models in multivariate analysis. In *Multivariate Analysis, Design of Experiments and Survey Sampling*(S. Ghosh, ed.)187-229. Dekker, New York.
- [5] Atay-Kayis, A. e Massam, H. (2005). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika*, to appear.
- [6] Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc.*, **25**, 220-223.
- [7] Bishop, Y. M., Fiernberg, S. e Holland, P. (1975). *Discrete Multivariate Analysis*. M.I.T. Press: Cambridge, Mass.
- [8] Csiszar, I. (1975). I-divergence geometry of probability distributions and minimisation problems. *Ann. Prob.*, **3**, 1, 146-158.
- [9] Cowell, R. G., Dawid, A. P., Lauritzen, S. L. e Spiegelhalter, D. J. (1999). *Probabilistic Networks and Experts Systems*. Springer, New York.
- [10] Darroch, J. N., Lauritzen, S. L. e Speed, T. P. (1980). Markov fields an log linear interaction models for contingency tables. *Ann. Estat.*, **8**, 522-539.
- [11] Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application, *Biometrika*, **68**, 265-274.
- [12] Dawid, A. P. e Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.*, **21**, 1272-1317.
- [13] Dellaportas, P., Giudici, e Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhya, Series A* **65**, 45-55.
- [14] Dempster, A. P. (1972). Covariance selection. *Biometrics*,**28**, 157-175.

- [15] Diaconis, P. e Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, **7**, 269-281.
- [16] Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics*, **42**, 204-223.
- [17] Dobra, A., Hans, C., Jones, B., Nevins, J. R., West, M. (2003). Sparse Graphical Models for Exploring Gene Expression Data. www.samsi.info/TR/tr2003-07.pdf.
- [18] Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G. e West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**, 196-212.
- [19] Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer-Verlag New York, Inc;
- [20] Flores, M. J., Gamez, J. A. e Olesen, K. G. (2003). Incremental compilation of Bayesian networks. In *Proc. 19th Annual Conference on Uncertainty in Artificial Intelligence*, 233-240. Morgan Kaufmann, San Francisco.
- [21] Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. MIT Press.
- [22] Freedman, D., Pisani, R. e Purves, R. (1978). *Statistics*. New York: Norton.
- [23] Fronk, E. M. (2002). Model Selection for Dags via RJMCMC for Discrete and Mixed Case. *Sonderforschungsbereich*, **386**, paper 271.
- [24] Fronk, E. M. e Giudici, P. (2000). Markov Chain Monte Carlo Model Selection for DAG models. *Statistical Methods & Applications*, **13**, 259-273.
- [25] Geiger, H. e Heckerman, D. (1999). Parameter Prioris for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions. Technical report, Microsoft Research.
- [26] Giudici, P. (1994). Learning in graphical Gaussina models, in: J. M. Bernardo, J. Berger, A. Dawid, A. Smith (Eds.), *Bayesian Statistics 5*, Oxford University Press, 621-628.
- [27] Giudici, P (1996). Learning in graphical Gaussian models. In J. M. Bernado, J. O. Berger, A. P. Dawid, and A. M. Smith (Eds), *Bayesian Statistics*, **5**, pp. 621-628. Oxford University Press.
- [28] Giudici, P. e Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785-801.
- [29] Giudici, P. e Stanghellini, E. (1999). Bayesian Inference for Graphical Factor Analysis Models. *Psychometrika*, **66**, 577-592.
- [30] Giudici, P., Green, P. J. e Tarantola, C. (1999). Efficient Model Determination for Discrete Graphical Models. *Biometrika*. To appear.
- [31] Goodman, L. A. (1970). The multivariate analysis of qualitative data: interaction among multiple classifications. *J. Amer. Statist. Soc. B*, **31**, 486-498.

- [32] Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**, 711-732.
- [33] Grone, R., Johnson, C. R. , Sa, E. M. e Wolkowice, H. (1984). Positive definitive completions of partial hermitian matrices. *Linear algebra and its applications* **58**, 109-124.
- [34] Haberman, S. J. (1974). The Analysis of Frequency Data. Univ. Chicago Press: Chicago.
- [35] Hammersley, J. M. e Clifford, P. E. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.
- [36] Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal Of Machine Learning Research*, **1**, 49-75.
- [37] Hofmann, R. e Tresp, V. (1998). Nonlinear Markov networks for continuous variables, in: M. I. Jordan, M. J. Kearns, S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference, MIT press, pp. 521-527.
- [38] Huang, E., West, M. e Nevins, J. R. (2003). Gene expression profiles and predicting clinical characteristics of breast cancer. *Recent Progress in Hormone Research*, 55-73.
- [39] Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M. -H., Horng, C. -F, Bild, A., Iversen, E., Liao, M., Chen, C. -M., West, M., Nevins J. e Huang, A. (2003). Gene expression predictors of breast cancer outcomes, *Lancet*, **361**, 1592-1596.
- [40] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. e West, M. (2004). Archival version including Appendicies: Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science*, **4**, 388-400.
- [41] Lampagne, C. B. (1979). *An Evolution of the Women and Mathematics (WAM) Program and Associated Sex-Related Differences in the Teaching, Learning and Counseling of Mathematics*. Ed. D. Thesis, Columbia University Teachers College, USA.
- [42] Lauritzen, S. L. Graphical Models (1996). Clarendon Press, Oxford.
- [43] Lauritzen, S. L. e Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, **17**, 31-57.
- [44] Lee, S. (2006). Sparse Graphical Models for exploring gene expression data. Final Project for STAT689 (Computing for Bioinformatics). Statistics, Texas A&M University
- [45] Madigan, D. e Raftery, A. E. (1994). Model Selection and Accoutuing for Model Uncertainty in Graphical Models using Occam's Window. *Journal of the Amercian Statistical Association*, **89**, 1535-1546.
- [46] Mardia, K. V., Kent, J. T. e Bibby, J. M. (1979). Multivariate Analysis. Academic Press: London.

- [47] *Para entender o básico*. Acedido em 22,03,2007. URL: <http://www.geocities.com/~sabio/genoma/projetogenoma2.htm>.
- [48] Pazza, R. (2006). *Introdução à Genética*. Acedido em 22,03,2007. URL: http://www.biociencia.org/index.php?option=com_content&task=view&id=211&Itemid=71.
- [49] Roverato, A. (2002). Hyper-inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* 29, 391-411.
- [50] Simpson, C. H. (1951). The interpretation of interaction in contingency tables. *J. R. Stat. Soc. B.* **13**, 238-241.
- [51] Speed, T. P. e Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 1, 184-197
- [52] Thulasiraman, K. e Swamy, M. N. S. (1992). *Graphs: Theory and Algorithms*. Wiley, New York.
- [53] Upton, G. J. G. (1991). The Exploratory Analysis of Suvey Data Using Log-Linear Models. *The Statistician*, **40**, 169-182.
- [54] Veer, L., Dai, H., Vijver, M., He, H., Hart, A., Mao, M., Peterse, H., Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R. e Friend, S. (2001). Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*, **415**, 530-536.
- [55] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.
- [56] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., R., M. J. e Nevins, J. R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of the Nationla Academy of Sciences*, **28**, 11462-11467.
- [57] Whittaker, J. (1990). *Graphical Models in Aplplied Multivariate Statistics*. John Wiley e Sons.
- [58] Wong, K. and Carter, C. (2002). An efficient sampler for decomposable covariance selection models. preprint.
- [59] Wong, F., Carter, C. e Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, **90**, 809-830.
- [60] Wright, S. (1921). Correlation and causation. *J. Agric. Res.*, **20**, 557-585.