



COPPE/UFRJ

MODELOS DE GRAFOS EM ESTATISTICA

Plácido Mateus Jequessene

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção.

Orientadores: Basílio de Bragança Pereira

Samuel Jurkiewicz

Rio de Janeiro

Março de 2010

MODELOS DE GRAFOS EM ESTATISTICA

Plácido Mateus Jequessene

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE PRODUÇÃO.

Examinada por:

Prof. Basílio de Bragança Pereira, Ph. D.

Prof. Samuel Jurkiewicz, D.Sc.

Prof^a. Laura Silvia Bahiense da Silva Leite, D.Sc.

Prof^a. Flávia Maria Pinto Ferreira Landim, D.Sc.

RIO DE JANEIRO, RJ – BRASIL.

MARÇO DE 2010

Jequessene, Plácido Mateus

Modelos de Grafos em Estatística/ Plácido Mateus
Jequessene – Rio de Janeiro: UFRJ/COPPE, 2010.

XIV, 97 p.: il.; 29,7 cm.

Orientadores: Basílio de Bragança Pereira e Samuel
Jurkiewicz

Dissertação (Mestrado) – UFRJ/ COPPE/ Programa de
Engenharia de Produção, 2010.

Referencias Bibliográficas: p. 95 - 97.

1. Modelos Grafos. 2. Independência Condicional. 3.
Análise Multivariada. I. Pereira, Basílio de Bragança *et al.*
II. Universidade Federal do Rio de Janeiro, COPPE,
Programa de Engenharia de Produção. III. Título.

Dedico aos meus pais Mateus Jequessene e Maria de Lurdes Augustinho Saize e
toda minha Família,
fontes de minha inspiração pela vida.

AGRADECIMENTOS

Muitos foram os que me ajudaram e torceram por mim para conquista de mais uma vitória na minha vida. Agradeço de todo coração o conforto, carinho, atenção dada a mim sem a preocupação de algo em troca.

Mas não citá-los aqui neste espaço seria ingratidão minha, mas acredito que posso perder a lembrança de alguém e por isso, desde já, peço desculpa por um eventual esquecimento.

O meu agradecimento primário vai ao senhor nosso Deus pela sua presença constante, pelo seu consolo que através da fé, consegue fortalecer minha rotina diária e dar-me força para enfrentar as dificuldades e alcançar os momentos felizes.

Dentre estas pessoas estão meus pais, duas riquezas inestimáveis na minha vida. É por causa deles que venho realizando meus sonhos e vontades, sempre cruzando batalhas em busca do meu sucesso. A eles, muito obrigado.

Quero agradecer a pessoa admirável, um verdadeiro Deus na terra, o Professor Basílio de Bragança Pereira meu orientador, pelo todo seu apoio desde que eu ainda me encontrava em Moçambique (País sudeste Africano), mesmo sendo eu um jovem desconhecido, confiou e tornou possível a minha candidatura ao curso. Ainda sou grato por vários motivos, pela confiança de orientar um jovem desconsolado, por mostrar o caminho do conhecimento e sabedoria, por contribuir para minha maturidade profissional, sentimental e pessoal. De todo coração vai para ele um meu Muito Obrigado.

Meus sinceros agradecimentos ao co-orientador, professor Samuel Jurkiewicz, por acreditar em mim e estar sempre presente, ajudando-me a solucionar problemas ao longo do percurso, pelas sugestões, os esclarecimentos e comentários oportunos e a compreensão silenciosa dos momentos difíceis pelos quais passei com o andamento deste trabalho.

Também agradeço a todos meus familiares em Moçambique que sentem muito minha falta, estão constantemente torcendo e orando por mim, citar nomes de todos seria exaustiva, a família é grande. Em especial os meus irmãos e aos meus Primos António Almeida Saize e Lino Alfredo que foram minha força motriz.

Arlinda Eliseu, cunhada do coração, muito obrigada pela presteza e apoio moral, psicológico que mesmo estando distante conseguia fazer chegar a mim. Você e meus sobrinhos Deisy e Marlon também foram muito importantes nesta minha trajetória.

Claro que não posso deixar de lembrar James Dean Oliveira dos Santos Júnior, um cara que até agora desempenha o papel de um irmão, que mesmo sem me conhecer direto, foi ele que nos meus primeiros dias no Rio, ajudou a encontrar um alojamento para minha estadia na Ilha do Governador, indicou caminho para COPPE-UFRJ e a obter toda documentação necessária para um estrangeiro no Brasil. Portanto, ao James o meu muitíssimo obrigado.

Agradeço a todos os colegas do mestrado, em especial aos que me marcaram e que nunca esquecerei porque estarão sempre no meu coração, muito obrigado.

Agradeço também a toda direção, professores e todos os funcionários do Curso de engenharia de produção na área de Pesquisa Operacional (COPPE-UFRJ), que sempre me ajudaram como puderam. Acredite que serei grato a cada funcionário desta brilhante Instituição. Deus lhes de a bênção.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

TEORIA DE GRAFOS E ESTATÍSTICA

Plácido Mateus Jequessene

Fevereiro de 2010

Orientadores: Basílio de Bragança Pereira

Samuel Jurkiewicz

Programa: Engenharia de Produção

Neste trabalho temos como objetivo ilustrar algumas aplicações da teoria de grafos em estatística, utilizando para isto os grafos, que auxiliam na visualização de estruturas de vários modelos estatísticos, portanto, os grafos representam a relação entre variáveis aleatórias, em que cada vértice é uma variável aleatória e as arestas entre os vértices representam as diferentes formas de relação entre as variáveis.

Nesta dissertação, através de grafos, foram visualizados estruturas de dependência e independência condicional, relações de causalidade entre variáveis aleatórias sob ponto de vista de redes Bayesianas. Apresentamos também grafos para vários modelos estatísticos aplicados em análise multivariada, como modelos de análise de regressão em situações de análise de regressão simples, regressão múltipla, regressão multivariada, equações estruturais, modelo log-linear e análise de séries temporais. Os modelos grafos aqui apresentado são modelos Gaussianos aonde se assume que as variáveis em análise ou em estudo são normalmente distribuídas.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

GRAPH THEORY AND STATISTICS

Plácido Mateus Jequessene

March/ 2010

Advisors: Basílio de Bragança Pereira

Samuel Jurkiewicz

Department: Industrial Engineering

The aim of this work is to illustrate some applications of graph theory in statistics, using different graphs, which help in the visualization of various statistical models' structures. Graphs represent the relationship between random variables, where each vertex is a random variable and the edges between the vertices represent the different forms of relationship between variables.

In this dissertation through graphs were visualized structures of dependence and conditional independence, relations of causality between random variables in terms of Bayesian networks. Also, graphs are presented for various statistical models applied in multivariate analysis, as a models of regression analysis in situations of simple regression analysis, multiple regression, multivariate regression, structural equation, log-linear model and time-series analysis. The models presented here are Gaussian graphical models, where it is assumed that the variables in the analysis or study are normally distributed.

SUMÁRIO

CAPÍTULO 1.....	1
INTRODUÇÃO	1
CAPÍTULO 2	5
FUNDAMENTOS DA TEORIA DOS GRAFOS	5
2.1. CONCEITOS BÁSICOS DE GRAFOS.....	5
2.1.1. Grafo não orientado.....	6
2.1.2. Grafo Orientado (ou Digrafo).....	10
2.2. GRAFOS ORIENTADOS ACÍCLICOS (DAG).....	19
2.2.1. Definição.....	19
2.2.2. Grafos de Cadeia (Chain Graph).....	20
2.2.3. Grafo Moralizado (Moral Graph).....	24
CAPÍTULO 3.....	26
CONCEITOS PROBABILÍSTICOS.....	26
3.1. INDEPENDÊNCIA CONDICIONAL EM GRAFOS	26
3.1.1. Propriedades de Markov.....	28
3.1.2. Propriedades de Markov em DAGs.....	31
3.1.3. Propriedade de Markov em Grafos de Cadeia.....	35
CAPÍTULO 4.....	39
REDES BAYESIANAS.....	39
4.1. CAUSALIDADE.....	39
4.2. REDES BAYESIANAS	43
4.2.1. Inferência em Redes Bayesianas.....	47
4.3. DIAGRAMA DE INFLUÊNCIA.....	51
CAPÍTULO 5.....	56
GRAFOS DE MODELOS ESTATÍSTICOS.....	56
5.1. ANÁLISE MULTIVARIADA.....	57
5.1.1. Variáveis Discretas – Modelo Log linear.....	57
5.1.2. Variáveis Contínuas – Modelo Gaussiano.....	61
5.2. ANÁLISE DE REGRESSÃO	68
5.2.1. Modelo de regressão Linear simples.....	69
5.2.2. Regressão Múltipla.....	71

5.2.3. <i>Regressão Multivariada</i>	72
5.3. EQUAÇÕES ESTRUTURAIS (OU EQUAÇÕES SIMULTÂNEAS).....	73
5.4. ANÁLISE DE SERIES TEMPORAL.....	80
5.4.1. <i>Modelos de Média Móveis</i>	82
5.4.2. <i>Modelo Auto-regressivo</i>	84
5.4.3. <i>Modelo Misto (Média Móvel e Auto-regressivo “ARMA”)</i>	85
5.4.4. <i>Modelos de Função de Transferência</i>	86
5.5. MODELOS EM ESPAÇO DE ESTADOS	89
5.6. MODELOS MECÂNICOS EM DINÂMICAS ECONÔMICAS	90
CAPÍTULO 6	93
CONSIDERAÇÕES FINAIS	93
Revisão Bibliográfica	95

LISTA DE TABELAS

Tabela 2.1: matriz de incidência.....	7
Tabela 2.2: matriz de adjacência.....	8
Tabela 4.1: valores de probabilidade dos eventos.....	48
Tabela 4.2: Custos em unidades monetárias (u.m).....	53
Tabela 5.1: Dados de comportamento dos ninhos de duas espécies de lagartos.....	60
Tabela 5.2: Matriz de variância de 5 disciplinas.....	64
Tabela 5.3: Matriz de correlações de 5 disciplinas.....	64
Tabela 5.4: Matriz inversa de correlações das 5 disciplinas.....	65
Tabela 5.5: Matriz inversa de correlações escalonada das 5 disciplinas.....	65
Tabela 5.6: Estrutura da matriz inversa de correlações escalonada.....	66
Tabela 5.7: matriz de correlação para valores observados, fonte, Dollin (1984).....	77

LISTA DE FIGURAS

Figura 1.1: ilustração das sete pontes sobre o rio Pregel.....	2
Figura 1.2: ilustração das sete pontes sobre o rio Pregel.....	2
Figura 2.1.1: ilustração de dois grafo.....	5
Figura 2.1.2: ilustração de subgrafo induzido.....	6
Figura 2.1.3: ilustração de grafo $G(4, 5)$	7
Figura 2.1.4: grafo completo K_4	9
Figura 2.1.5: ilustração de um Clique.....	9
Figura 2.1.6: Grafos bipartidos.....	10
Figura 2.1.7: Grafos orientados.....	10
Figura 2.1.8: exemplo de um grafo orientado.....	11
Figura 2.1.9: ilustração de uma cadeia e caminho em grafo.....	13
Figura 2.1.10: ilustração de grafo de Euler.....	14
Figura 2.1.11: grafo ilustrando fecho transitivo.....	15
Figura 2.1.12: grafo conexo não orientado (a) e grafo orientado conexo (b).....	15
Figura 2.1.13: grafo não orientado desconexo.....	15
Figura 2.1.14: ilustração de grafo fracamente conexa.....	16
Figura 2.1.15: ilustração de árvore (a) e árvore binária (b).....	17
Figura 2.1.16: arborescência.....	17
Figura 2.1.17: Grafo valorado ou Rede.....	18
Figura 2.1.18: Multigrafo.....	19
Figura 2.2.1: grafo direcionado acíclico (a) e direcionado não acíclico (b).....	19
Figura 2.2.2: grafo de cadeia (a) e grafo não de cadeia (b).....	21
Figura 2.2.2c: grafo de cadeia.....	22
Figura 2.2.2d: grafo de cadeia.....	22
Figura 2.2.2e: grafo de cadeia desencadeando em bloco.....	23
Figura 2.2.3: grafo orientados acíclico (a, c) e o seu grafo moralizado (b, d).....	24
Figura 3.1: grafo ilustrando a relação de dependência de quatro variáveis.....	27
Figura 3.1.1a. Propriedade de Markov para pares de variáveis.....	29
Figura 3.1.1b. Propriedade de Markov Local.....	30
Figura 3.1.1c. Propriedade de Markov Global.....	30

Figura 3.1.2a: grafo orientado acíclico (DAG) simples.....	31
Figura 3.1.2b, c: grafo orientado acíclico (b) e o seu grafo moralizado (c).....	32
Figura 3.1.2d, e: dois grafos orientados acíclicos ilustrando o caminho entre V e W....	34
Figura 3.1.3a: grafos de cadeia.....	37
Figura 3.1.3b: DAG dos componentes da cadeia em ‘a’.....	37
Figura 3.1.3c: Um grafo orientado mostrando que G é independente de R.....	38
Figura 4.1a, b, c: tipos de conexões numa rede causal.....	40
Figura 4.1d: grafo representando uma rede de casualidade.....	42
Figura 4.2a: grafo orientado acíclico (DAG).....	44
Figura 4.2b: grafo ilustrando teste de gravidez.....	45
Figura 4.2c: Grafo ilustrando teste de gravidez com um vértice auxiliar.....	46
Figura 4.2d: exemplo de rede bayesiana com muitos pais.....	46
Figura 4.2e: rede bayesiana depois da separação de pais.....	47
Figura 4.2.1: exemplo de rede bayesiana.....	47
Figura 4.3a: diagrama de influência do exemplo 1.....	53
Figura 4.3b: árvore de decisão do exemplo 1.....	54
Figura 4.3c: diagrama de influência do exemplo 2.....	55
Figura 5.1.1a: modelo grafo para variáveis A, B e C independentes.....	56
Figura 5.1.1b: modelo grafo para variáveis A, B e C onde $B \perp\!\!\!\perp C \mid A$	58
Figura 5.1.1c: modelo grafo não hierárquico AB, BC e AC.....	60
Figura 5.1.1d: grafo resultante no modelo AB, AC.....	61
Figura 5.1.2a: grafo não orientado com quatro vértices.....	63
Figura 5.1.2b: grafo de independência condicional de cinco disciplinas.....	66
Figura 5.1.2c: grafo ilustrando a propriedade global de Markov em cinco disciplinas..	68
Figura 5.2.1: grafo para regressão linear simples.....	69
Figura 5.2.2a: grafo para regressão linear múltipla.....	71
Figura 5.2.2b: grafo para regressão linear múltipla com três variáveis explicativas.....	72
Figura 5.2.3: grafo para regressão linear multivariada.....	73
Figura 5.3a: grafo para equações estruturais ou simultâneas.....	74
Figura 5.3b: diagrama de caminhos para o desempenho e a satisfação.....	76
Figura 5.3c: grafo mostrando a relação entre três variáveis independentes e duas dependentes.....	78

Figura 5.3d: grafo com os coeficientes dos efeitos entres as variáveis no modelo de Dollin (1984).....	79
Figura 5.4: modelo gráfico de três series temporais.....	82
Figura 5.4.1: grafo para uma estrutura de um modelo de média móvel de ordem três...	83
Figura 5.4.2: grafo para uma estrutura de modelo auto-regressivo de ordem três.....	84
Figura 5.4.3: grafo para uma estrutura de modelo ARMA (1, 2).....	85
Figura 5.4.4a: Grafo de um modelo de variáveis defasada L(3).....	86
Figura 5.4.4b: Grafo de um modelo de função de transferência L(3) + ARMA.....	87
Figura 5.4.4c: Grafo de um modelo de diferença estocástica D(2, 3).....	88
Figura 5.4.4d: Grafo de um modelo de diferença estocástica D(2, 3) + ARMA.....	88
Figura 5.5: grafo do modelo espaço de estado.....	89
Figura 5.6: grafo representando um modelo simples de Phillips.....	92

Capítulo 1

Introdução

O presente trabalho é de caráter didático cujo intuito principal é desenvolver um material para o ensino e aprendizagem que integre fundamentos de teoria de grafos e sua modelagem em estatística, sem, no entanto, apresentação de algoritmos para resolução dos diversos modelos estatísticos aqui apresentados.

A Teoria de Grafos está presente no nosso cotidiano, desde as redes de estradas das cidades até o mais simples problema de colorir, por exemplo, mapas. Ela tem aplicação em muitas áreas do conhecimento, tais como: Física, Química, Biologia, Computação, Pesquisa Operacional, Estatística, Engenharia, Psicologia, Sociologia, Genética, Telecomunicações, Escolha de uma rota ótima e Logística, dentre outros.

A coloração de mapas é um dos problemas clássicos no qual é utilizada a teoria de grafos. Quantas cores serão necessárias para colorir um mapa, de forma que todos os países tenham cores diferentes dos que lhe fazem fronteira? A resposta é quatro, no máximo, qualquer que seja o mapa.

Por exemplo, as eliminatórias de um torneio esportivo podem ser esquematizadas usando um grafo. Esta teoria, em constante expansão, ajuda a resolver problemas nas mais variadas áreas, e a tendência é para se desenvolver mais e mais.

Ao contrário de muitos ramos da matemática, nascidos de especulações puramente teóricas, a teoria dos grafos tem 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.

O pai desta teoria é **Leonhard Paul Euler** (1707-1783), o célebre matemático que se evidenciou em muitas outras áreas científicas. Euler levantou uma questão que se

relacionava com a arquitetura e com o ordenamento urbano da cidade de Königsberg (Prússia até 1945, atual Kaliningrado, Rússia) que fica em uma pequena porção da Rússia, entre a Polônia e a Lituânia, no continente europeu.

A cidade era cruzada pelo Rio Pregel, e tinha duas pequenas ilhas centrais. Uma das ilhas era ligada a cada margem por duas pontes. A outra tinha duas pontes, cada uma ligando-a a uma margem. Existia ainda uma sétima ponte ligando as duas ilhas.

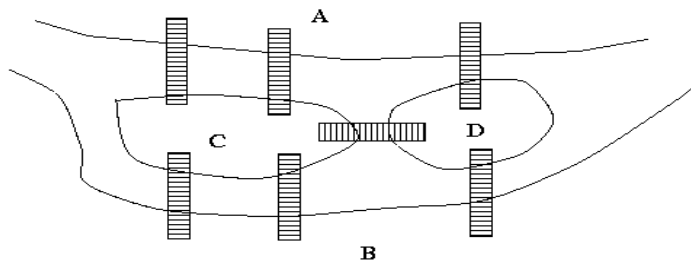


Figura 1.1: ilustração das sete pontes sobre o rio Pregel

A pergunta era: seria possível iniciar o percurso numa das quatro zonas (à esquerda das duas ilhas, à direita das duas ilhas, em cada uma das margens), e percorrer todas as pontes sem repetir nenhuma?

Euler desenhou um diagrama, atribuindo um vértice a cada uma das áreas, e uma linha a cada ponte. Assumindo cada zona como distinta e atribuindo a cada uma delas uma partida e uma chegada, então, se apenas se passar por cada zona e por cada ponte uma vez, apenas haverá uma rota possível, isto se as ligações entre as quatro zonas forem constantes. Euler provou mais tarde que, neste caso, não havia solução.

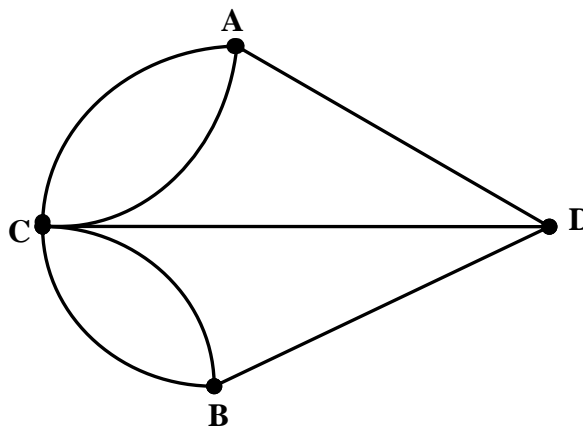


Figura 1.2: ilustração das sete pontes sobre o rio Pregel

Esta questão, na altura relacionada com um simples problema de melhorar o quotidiano aos habitantes da cidade, facilitando-lhes o seu deslocamento de um lado a outro do rio, foi o começo da Teoria dos Grafos.

Apesar disto, quando Euler propôs o problema, este foi considerado desinteressante e sem aplicações práticas, e caiu no esquecimento durante um século, até ser novamente aproveitado na Química por Cayley, na Biologia por Jordan e na Engenharia Elétrica por Kirchoff.

Muitas vezes, para resolver uma determinada situação problemática, temos tendência de fazer um esquema, ou um modelo, que nos facilite a organização das idéias. Com base nesses modelos, conseguimos visualizar melhor qual é a solução para o nosso problema ou definir uma estratégia para resolvê-lo.

Em muitas situações os tipos de modelos utilizados, são grafos, que não são mais do que esquemas nos quais se utilizam pontos ligados por linhas conforme a relação que é estabelecida no problema.

Os grafos são freqüentemente utilizados para visualizar melhor muitos problemas analisados em estatística (entre outros, as redes Bayesianas, a análise de regressão...), facilitando, deste modo, a definição de uma estratégia para resolvê-los.

Um estudo retrospectivo da literatura nos leva diretamente aos trabalhos do geneticista Sewall Wright, surgidos nos idos de 1918, 1920 e 1921, nos quais o autor declara ter encontrado um método flexível de relacionar, num sistema multidimensional, os coeficientes de correlação entre variáveis às relações funcionais existentes entre elas.

Foi com o aparecimento do computador que este assunto realmente ganhou notoriedade, e os cientistas se aperceberam da enorme utilidade que poderia ter na resolução do mais variado tipo de questões.

Hoje em dia, com programas complexos da tecnologia computacional, criaram-se inúmeros modelos e algoritmos de resolução, que, baseados na teoria dos grafos,

sugerem automaticamente qual a melhor maneira de disposição, qual a solução ótima para o problema, etc.

Os grafos podem ser usados para visualizar informação relacionada com a estrutura de independência condicional existente entre as variáveis ou 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.

Uma das principais características de modelação através de grafos na estatística é a simplificação, onde qualquer procedimento sistemático para analisar observações multivariadas deverá condensar o conjunto de dados sem eliminar ou obscurecer as associações relevantes existentes.

Os modelos grafos fornecem uma estrutura unificada para análise estatística de dados contínuos, resumidos pela matriz de correlações ou covariâncias, e dados discretos resumidos por uma tabela de contingência, e essa unificação sugere generalizações para sistemas de variáveis mistas.

Encontramos nos grafos um forte apelo para a representação de redes probabilísticas, estruturas de inter-relacionamento e dependência entre variáveis. Os grafos são muito eficientes na modelagem de muitos problemas.

Para se alcançar os propósitos deste trabalho serão primeiramente definidos alguns conceitos ligados a teoria de grafos e posteriormente através de exemplos, alguns modelos estatísticos serão modelado a partir de grafo.

Capítulo 2

Fundamentos da teoria dos grafos

Neste capítulo, serão apresentados e definidos alguns conceitos básicos e outros poucos citados em muitos livros de teoria de grafos, entre eles o caso de grafos Moralizados, dos quais, alguns serão uteis para o desenvolvimento ou descrição das aplicações de grafos em estatística nos capítulos mais adiante deste trabalho.

2.1. Conceitos básicos de grafos

O que é um grafo?

Um grafo pode ser descrito como uma estrutura constituída por dois elementos fundamentais: os arcos ou arestas e os vértices. Cada arco tem uma origem (ponta inicial) e um destino (ponta final), quase como uma estrada, que sai de uma cidade e chega a outra. Neste caso os vértices são as cidades.

Em outras palavras, grafo pode ser considerado como sendo uma noção abstrata e intuitiva, usada para representar a idéia de alguma espécie de relação entre os “objetos”. Graficamente, aparece representado por uma figura com *nós* ou *vértices*, significando os objetos, unidos por um traço denominado *aresta* configurando a relação imaginada.

Matematicamente chama-se grafo a um par $G = (V, A)$, tal que $V = V(G) = \{v_1, \dots, v_n\}$ é o conjunto dos vértices (não vazio e finito) e $A = A(G)$ é o conjunto das arestas ou ligações entre os vértices, isto é, $A(G) = \{a_1, \dots, a_m\}$, com $a_k = \{v_{ki}, v_{kj}\}$, para $k \in \{1, \dots, m\}$, ($|V| = n$, $|A| = m$).

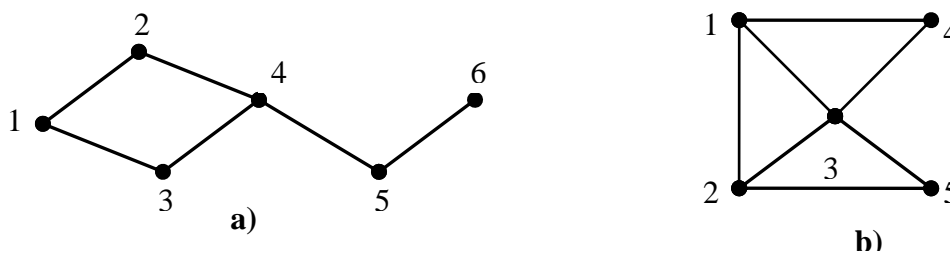


Figura 2.1.1: ilustração de dois grafo

As figuras acima mostram dois grafos $G=(V,A)$, nos quais na figura 2.1.1a, $V=\{1,2,3,4,5,6\}$ e $A=\{\{1,2\},\{1,3\},\{2,4\},\{3,4\},\{4,5\}$ e $\{5,6\}\}$ ($|V|= 6$, $|A|= 6$) e na figura 2.1.1b, $V=\{1,2,3,4,5\}$ e $A=\{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,5\}\{3,4\},\{3,5\}\}$; $|V|=5$, $|A|=7$. É comum a utilização da variável v_i ou x_i , $i=1, 2, \dots, n$ para a distinção dos nós (vértices).

2.1.1. Grafo não orientado

Um grafo não orientado é um par $G=(V, A)$ em que V é um conjunto finito de vértices e A de arestas, no qual as arestas não são pares ordenados de vértices, isto é, uma aresta $(v; w)$ é igual a $(w; v)$, exemplo figura 2.1.1a e 2.1.1b.

Dois vértices v e w são **adjacentes** (ou vizinhos) se há uma aresta $a = (v, w)$ em G . Esta aresta é dita ser incidente a ambos (v e w). Por exemplo, os vértices 2 e 4 em figura 2.1.1a.

Portanto, a vizinhança do vértice $N(v)$ é o conjunto dos vértices adjacentes a v , isto é:

$$N(v) = \{w \mid (v, w) \in A\}$$

Em um grafo não orientado grafo G , designa-se por **grau** ou valência de um vértice “ v ” (denota-se por $d(v)$, ou grau(v)) o número de arestas de G incidentes em “ v ”. Na figura 2.1.3, o vértice três (v_3) tem grau três.

Um **subgrafo** $G'(V', A')$ de um grafo $G(V, A)$ é um grafo em que $V' \subset V$ e $A' \subset A$.

Subgrafo induzido por um subconjunto $W \subset V$ é o grafo que tem W como conjunto de vertices e cujas arestas são as arestas de A que ligam os vertices de W .

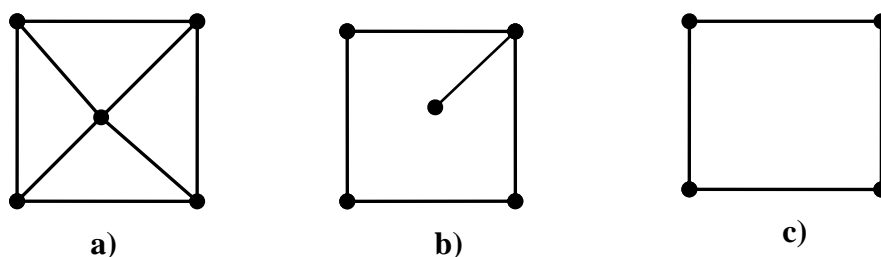


Figura 2.1.2: ilustração de subgrafo induzido (c)

Dos grafos a), b) e c) cujos diagramas são dados na figura 2.1.2, podemos dizer que (b) é um subgrafo induzido de (a) enquanto que (c) é um subgrafo, mas não é induzido de (a).

Matriz de incidência é uma representação de um grafo através de uma matriz bidimensional, que guarda informações sobre como os vértices se relacionam com cada aresta (isto é, informações sobre a *incidência* de um vértice em uma aresta), portanto, uma das dimensões consiste de vértices e a outra dimensão consiste em arestas.

Para representar um grafo sem pesos nas arestas e não direcionado, basta que as entradas da matriz M ($n \times m$) contêm 1 se o vértice incide na aresta, 2 caso seja um laço (incide duas vezes) e 0 caso o vértice não incida na aresta.

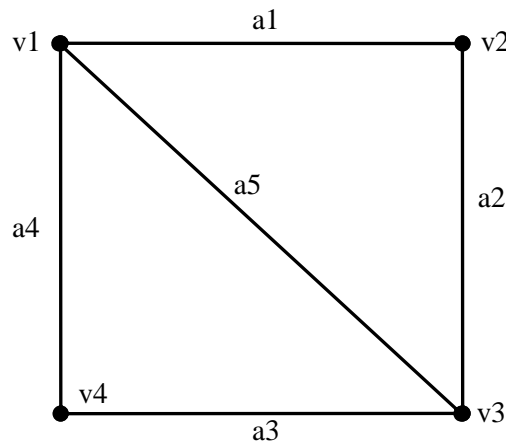


Figura 2.1.3: ilustração de grafo $G(4, 5)$

Por exemplo, a matriz de incidência do grafo acima (figura 2.1.3) é

Tabela 2.1: matriz de incidência

	a ₁	a ₂	a ₃	a ₄	a ₅
v ₁	1	0	0	1	1
v ₂	1	1	0	0	0
v ₃	0	1	1	0	1
v ₄	0	0	1	1	0

Matriz de adjacência é uma representação de um grafo através de uma matriz bidimensional, que guarda informações sobre como os vértices v_i e v_j estão relacionados (isto é, informações sobre a *adjacência* de v_i e v_j).

A matriz de adjacências nos dá informações relativas a uma forma limitada de alcançabilidade, as que são obtidas através de caminhos de comprimento 1.

Para representar um grafo não direcionado, simples e sem pesos nas arestas, basta que as entradas m_{ij} de uma matriz de adjacência $M = m_{ij}$ contenha 1 se v_i e v_j são adjacentes e 0 caso contrário. Por exemplo, a matriz de adjacência do grafo (figura 2.1.3) é

Tabela 2.2: matriz de adjacência

	v_1	v_2	v_3	v_4
v_1	0	1	1	1
v_2	1	0	1	0
v_3	1	1	0	1
v_4	1	0	1	0

Se as arestas do grafo tiverem pesos, m_{ij} pode conter, ao invés de 1 quando houver uma aresta entre v_i e v_j , o peso dessa mesma aresta.

Um **conjunto independente** de um grafo G é um conjunto S de vértices de G tal que não existem dois vértices adjacentes contidos em S . Em outras palavras, se v e w são vértices quaisquer de um conjunto independente, não há aresta entre v e w .

Grafo completo é o grafo (não orientado) no qual quaisquer dois vértices distintos são adjacentes, isto é, todo vértice é adjacente a todos os outros vértices (todos os vértices têm grau $n - 1$). Para cada $n \geq 1$, o grafo completo com n vértices é denotado por K_n .

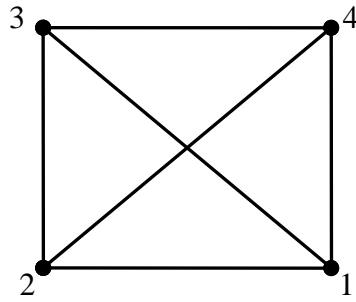


Figura 2.1.4: grafo completo K_4

Uma **clique** em um grafo G é um subgrafo de G que é completo. O tamanho de uma clique é igual a cardinalidade de seu conjunto de vértices. Por exemplo no grafo abaixo (figura 2.1.5) sendo “ V ” seu conjunto de vértices e “ A ” o de arestas, temos que:

Se $V=\{1,2,3,4,5\}$ e $A=\{(1,2),(1,3),(1,4),(1,5),(2,3), (2,4),(3,4),(4,5)\}$, o subgrafo induzido pelos vértices $(1,2,3,4)$ é uma clique de tamanho 4 e $(1,4,5)$ é uma clique de tamanho 3.

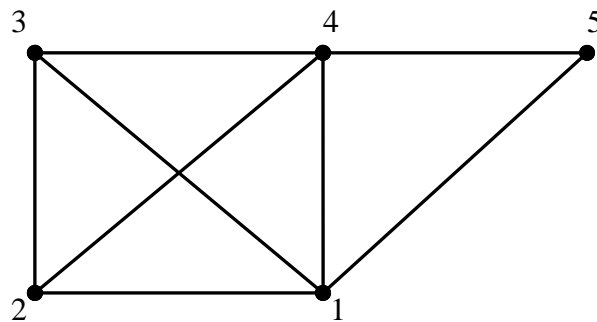


Figura 2.1.5: ilustração de um Clique

Um **grafo bipartido** é o grafo cujos vértices podem ser divididos em dois subconjuntos, digamos V_1 e V_2 , de tal modo que não há arestas entre vértices de um mesmo subconjunto, ou (dizendo de outra maneira) de modo que dois vértices do mesmo conjunto não sejam adjacentes (figura 2.1.6a e 2.1.6b).

Um **grafo bipartido completo** é um grafo bipartido, onde qualquer vértice do primeiro conjunto é adjacente a todos vértices do segundo conjunto. Eles são representados por $K_{r,s}$, onde r e s são os numeros de vértices dos dois conjuntos ($K_{3,2}$, figura 2.1.6b).

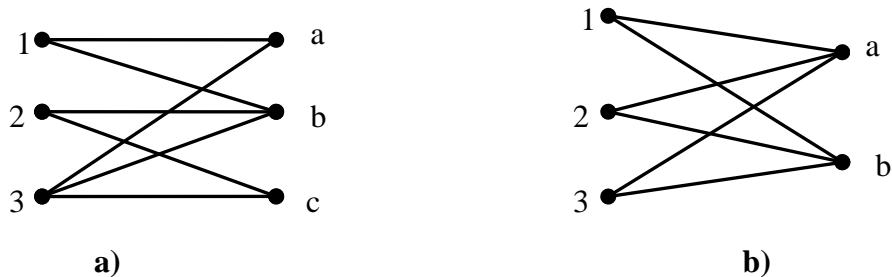


Figura 2.1.6: Grafos bipartidos

2.1.2. Grafo Orientado (ou Digrafo)

Um Grafo é dito grafo orientado, dirigido ou *Digrafo* quando o sentido das ligações entre os vértices é considerado. As arestas dos Grafos Orientados são denominadas "arcos" (setas) ou "arestas orientadas".

Uma aresta orientada é um par ordenado $(v_1; v_2)$ do conjunto "A" onde a ordem dos elementos v_1 e v_2 faz diferença: $(v_1;v_2)$ indica uma aresta orientada do vértice v_1 para o vértice v_2 , $(v_2;v_1)$ indica uma aresta orientada do vértice v_2 para o vértice v_1 .

Um Digrafo é também representado matematicamente por:

$G=(V,E)$, onde V é o conjunto de vértices e "A" é uma relação binária em "V" (i.e., um conjunto de *pares ordenados*) das ligações.

$A= \{ v_1,v_2 \in V, (v_1;v_2) \text{ diferente de } (v_2;v_1) \}$, entretanto $(v_i;v_j)$ pode-se notar por $v_i \rightarrow v_j$.

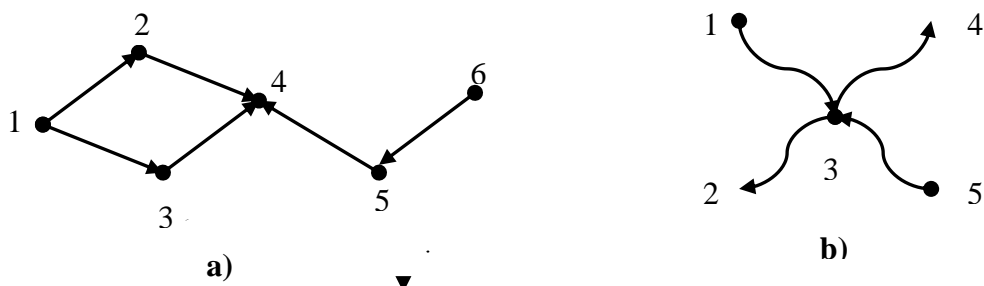


Figura 2.1.7: Grafos orientados

As figuras acima mostram dois grafo $G=(V,A)$ orientados ou direcionados, onde para a figura 2.1.7a, temos $V=\{1,2,3,4,5,6\}$ e $A=\{\{1,2\},\{1,3\},\{2,4\},\{3,4\},\{5,4\}, \{6,5\}\}$ e para a figura 2.1.7b, $V=\{1,2,3,4,5\}$ e $A=\{\{1,3\},\{3,2\},\{3,4\},\{5,3\}\}$; $|V|=5$, $|A|=4$.

Considere, agora, o grafo definido por:

$V = \{p \text{ dado que } p \text{ é uma pessoa da família}\}$

$A = \{(v,w) \text{ dado que } (v \text{ é pai ou mãe de } w) \}$

Um exemplo de deste grafo (ver Figura 2.1.8) abaixo:

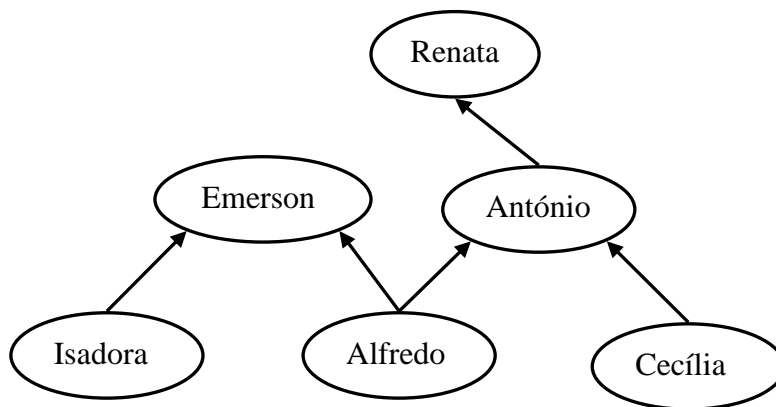


Figura 2.1.8: exemplo de um grafo orientado
 Fonte: <http://www.inf.ufsc.br/grafos/definicoes/definicao.html>

$V = \{Emerson, Isadora, Renata, Antonio, Rosane, Cecília, Alfredo\}$

$A = \{(Isadora, Emerson), (Alfredo, Emerson), (Alfredo, Antonio), (Cecília, Antonio), (Antonio, Renata)\}$

A relação definida por “A” não é simétrica, pois se “v” é pai ou mãe de “w”, não é o caso de “w” ser pai ou mãe de “v”. Há, portanto, uma orientação na relação, com um correspondente efeito na representação gráfica de G.

Em um grafo orientado (por exemplo, a figura 2.1.8), a adjacência (vizinhança) é especializada em:

Sucessor: um vértice w é sucessor de v se há um arco que parte de v e chega em w . Por exemplo, na figura 2.1.8, diz-se que *Emerson* e *Antonio* são sucessores de *Alfredo*.

Antecessor: um vértice v é antecessor de w se há um arco que parte de v e chega em w .

Na figura 2.1.8, por exemplo, diz-se que *Alfredo* e *Cecília* são antecessores de *Antonio*.

Vale salientar que o número de vértices em grafo orientado G é chamado sua *ordem* e o número de arcos em G o seu *tamanho*.

Num grafo orientado os graus podem ser de entrada ou de saída, a noção de grau é especializada em:

- **O grau de entrada** ou chegada ($d^-(v)$) de um vértice “ v ” é o número de arcos que lhe chegam ou que chegam em “ v ”. Para a figura 2.1.8, temos:
 - ✓ Grau de entrada ou chegada (*Antonio*) = 2
 - ✓ Grau de entrada ou chegada (*Alfredo*) = 0
 - ✓ Grau de entrada ou chegada (*Renata*)=1

- **O grau de saída** ou Emissão ($d^+(v)$) corresponde ao número de arcos que o têm como origem. Por exemplo, no grafo em 2.1.8, temos:
 - ✓ Grau de Emissão ou saída (*Antonio*) = 1
 - ✓ Grau de Emissão ou saída (*Alfredo*) = 2
 - ✓ Grau de Emissão ou saída (*Renata*) = 0

Um vértice v é uma **Fonte** se grau de Recepção(v) = 0. É o caso dos vértices *Isadora*, *Alfredo* e *Cecília*.

Um vértice v é um **Sumidouro** se grau de Emissão(v) = 0. É o caso dos vértices *Renata* e *Emerson*.

Um **Lazo** é uma aresta ou arco do tipo $a = (v, v)$, ou seja, que relaciona um vértice a ele próprio. Ou ainda, é uma aresta ou arco em que a extremidade inicial e final é coincidente.

Uma **cadeia** é uma seqüência qualquer de arestas adjacentes que ligam dois vértices. O conceito de cadeia vale também para grafos orientados, bastando que se ignore o sentido da orientação dos arcos. A seqüência de vértices (v_1, v_2, v_3, v_4) é um exemplo de cadeia (figura 2.1.9a).

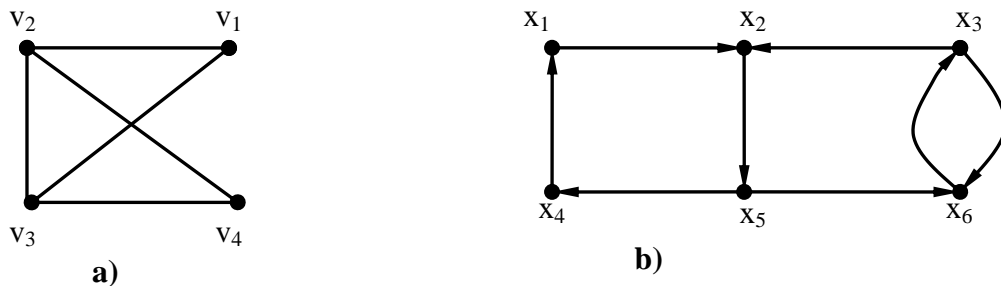


Figura 2.1.9: ilustração de uma cadeia e caminho em grafo

Uma cadeia é dita ser **elementar** se ao percorrermos o grafo não passamos duas vezes pelo mesmo vértice, ex: (v_1, v_3, v_2) na figura 2.1.9a.

É dita ser **simples** se ao percorrermos o grafo não passamos duas vezes pela mesma aresta (arco), ex: $(v_1, v_3, v_4, v_2, v_3)$ na figura 2.1.9a.

O **comprimento** de uma cadeia é o número de arestas (arcos) que a compõe.

Um **caminho** é uma cadeia na qual todos os arcos possuem a mesma orientação. Ou é uma seqüência de vértices adjacentes em que a extremidade final de uma aresta (arco) é extremidade inicial da aresta (arco) seguinte. A seqüência de vértices $(x_1, x_2, x_5, x_6, x_3)$ é um exemplo de caminho em 2.1.9b.

Um **ciclo** é uma cadeia simples e fechada (o vértice inicial é o mesmo que o vértice final). A seqüência de vértices $(x_1, x_2, x_5, x_4, x_1)$ é um exemplo de ciclo elementar em 2.1.9b.

Dentro dos ciclos existem:

Ciclo Euleriano é ciclo que passa uma vez por todas as arestas do grafo, sem repetir nenhum. É sinônimo de cadeia simples (percorrendo todas as arestas do grafo). Todos os grafos que admitem um caminho de Euler são chamados de grafos de Euler.

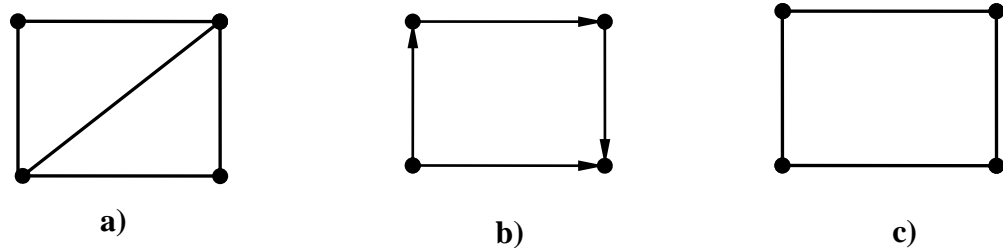


Figura 2.1.10: ilustração de grafo de Euler

Na figura 2.1.10, os grafo em “a” e “b” não são grafo de Euler, a penas o grafo em “c” é de Euler.

Ciclo Hamiltoniano é uma cadeia que começa e termina no mesmo vértice, percorrendo todos os vértices uma só vez (exceto o primeiro que também é o último).

Um **circuito** é um caminho simples e fechado. A seqüência de vértices $(x_1, x_2, x_5, x_4, x_1)$ é um exemplo de circuito elementar em 2.1.10b.

O **fecho transitivo direto (ftd)** de um vértice v é o conjunto de todos os vértices que podem ser atingidos por algum caminho iniciando em v . O *ftd* do vértice x_5 do grafo na figura 2.1.11, por exemplo, é o conjunto: $\{x_1, x_2, x_3, x_4, x_5, x_6\}$. Note que o próprio vértice faz parte do *ftd* já que ele é alcançável partindo-se dele mesmo.

O **fecho transitivo inverso (fti)** de um vértice v é o conjunto de todos os vértices a partir dos quais se pode atingir v por algum caminho. O *fti* do vértice x_5 do grafo na figura 2.1.11, por exemplo, é o conjunto: $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$. Note que o próprio vértice faz parte do *fti* já que dele se pode alcançar ele mesmo.

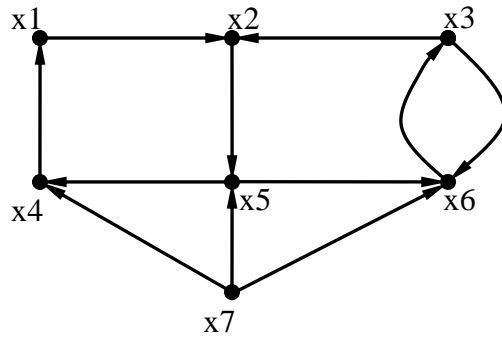


Figura 2.1.11: grafo ilustrando fecho transitivo

Um grafo $G(V, A)$ é dito ser **conexo** se há pelo menos uma cadeia ligando cada par dos seus vértices, isto é, se $G(V, A)$ é conexo, então $G(V, A)$ não pode admitir um conjunto de vértices particionado com dois subconjuntos tais que ambos os terminais de cada aresta pertençam ao mesmo subconjunto (figura 2.1.12a e 2.1.12b).

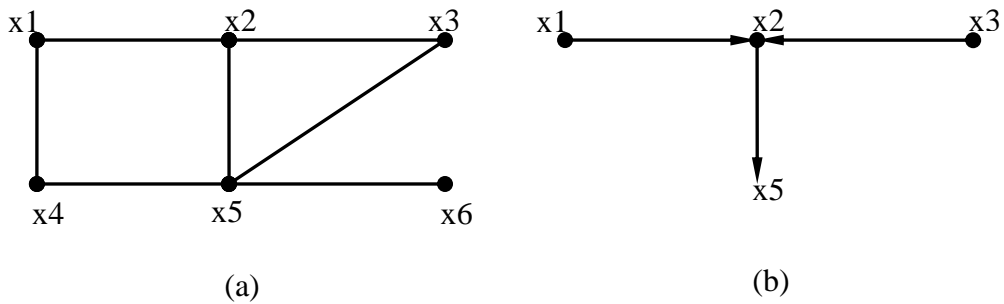


Figura 2.1.12: grafo conexo não orientado (a) e grafo orientado conexo (b)

Um grafo $G(V, A)$ é dito ser **desconexo** se há pelo menos um par de vértices que não está ligado por nenhuma cadeia (figura 2.1.13).

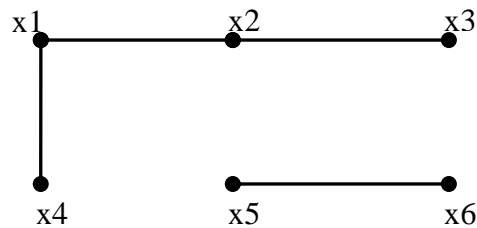


Figura 2.1.13: grafo não orientado desconexo

Todo grafo desconexo é composto por subgrafos conexos chamados de componentes. Por exemplo, o grafo 2.1.13 é um grafo desconexo composto por duas componentes.

Para caso de grafos orientados, um grafo é dito *ser fortemente conexo (f-conexo)* se todo par de vértices está ligado por pelo menos um caminho em cada sentido, ou seja, se cada par de vértices participa de um circuito. Isto significa que cada vértice pode ser alcançado partindo de qualquer outro vértice do grafo (figura 2.1.9b, 2.1.14a).

Conectividade **fracamente conexa** é considerada para grafo orientado tal que, ignorando o sentido dos arcos, para todo o vértice v, w , existe um caminho de v para w (figura 2.1.12b, 2.1.14b).

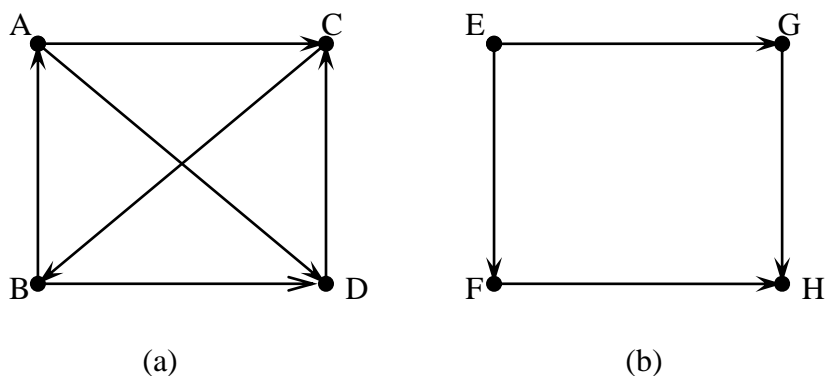


Figura 2.1.14: ilustração de grafo fracamente conexa

Uma **árvore** é um grafo não orientado conexo sem ciclos, logo não origina um circuito.

Seja $G(V, A)$ um grafo com ordem $n > 2$; as propriedades seguintes são equivalentes para caracterizar G como uma árvore:

1. G é conexo e sem ciclos;
2. G é sem ciclos e tem $n-1$ arestas;
3. G é conexo e tem $n-1$ arestas;
4. G é sem ciclos e por adição de uma aresta se cria um ciclo e somente um;
5. G é conexo, mas deixa de sê-lo se uma aresta é suprimida (todas as arestas são pontes);

6. Todo par de vértices de G é unido por uma e somente uma cadeia simples.

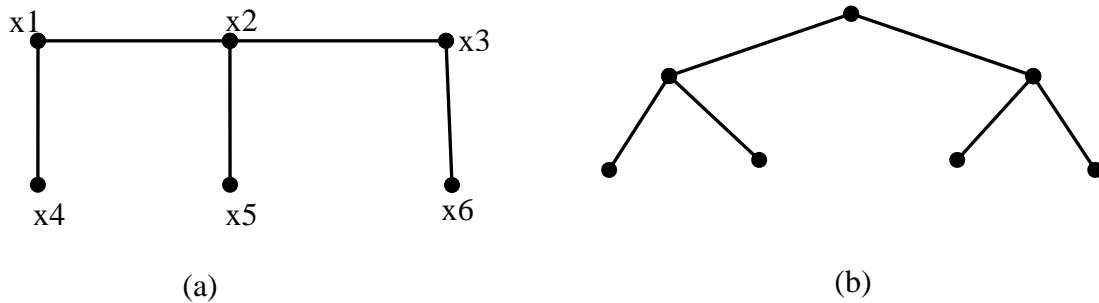


Figura 2.1.15: ilustração de árvore (a) e árvore binária (b)

Um tipo de estrutura associado a árvores é a arborescência. Usa-se este termo para diferenciá-las das árvores em que não consideramos uma orientação. Na arborescência, escolhemos um vértice na árvore como “raiz” da árvore e teremos uma orientação “natural” dos arcos.

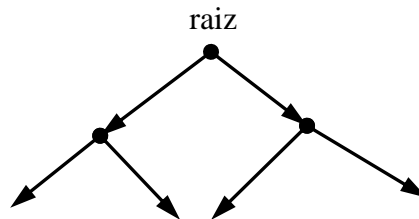


Figura 2.1.16: arborescência

Esta característica nos autoriza a não indicar as setas, se assim preferir: a raiz dá o sentido das setas. Portanto, em uma arborescência qualquer, existe um vértice sem antecessores (a raiz) e todos os vértices (fora a raiz) possuem exatamente um único antecessor.

Em muitas situações usa-se frequentemente uma arborescência binária (usualmente chamada de árvore binária), a qual tem uma particularidade adicional de que todos os vértices têm no máximo dois sucessores (por ser binária) ou cada vértice não tem grau maior que dois.

Grafos valorados (Redes ou Networks)

Uma Rede é um grafo orientado (ou um dígrafo) no qual um número real é associado aos vértices e/ou ligações. Este número é freqüentemente referido como o *peso* da ligação. Essa classificação é dada de acordo com a necessidade, ou não, da indicação do fluxo entre os vértices.

Na prática este número pode representar:

- custos, distâncias, capacidades, e/ou suprimentos e demandas;
- tempo (trânsito, permanência, etc.);
- confiabilidade de transmissão;
- probabilidade de ocorrerem falhas;
- capacidade de carga; e outros.

Uma Rede matematicamente também é representada por:

$G=(V,A,w)$ Onde: V é o conjunto de vértices; A é o conjunto de ligações (arestas ou arcos); e w é o peso associado aos vértices e/ou ligações.

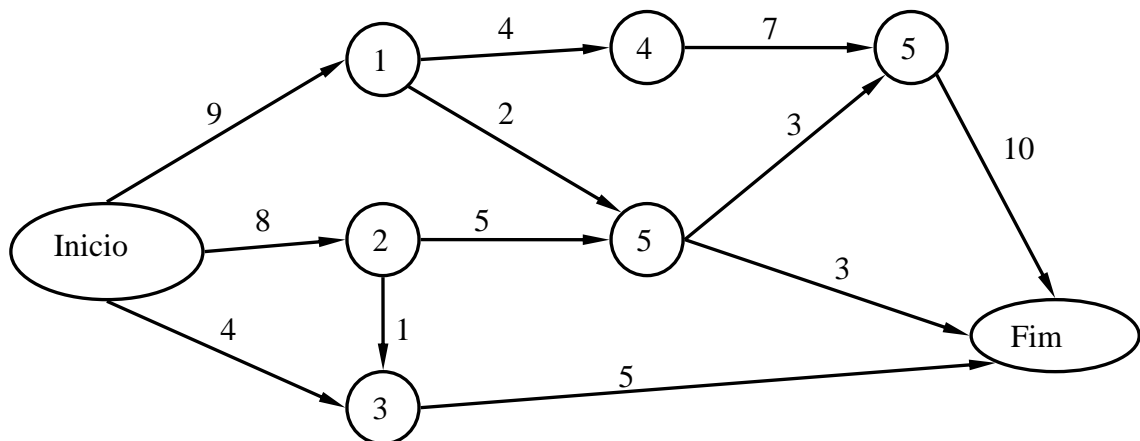


Figura 2.1.17: Grafo valorado ou Rede

Um grafo $G(V,A)$ é dito ser um **Multigrafo** quando existem múltiplas arestas entre pares de vértices de G . No grafo abaixo da figura 2.1.18, por exemplo, há duas arestas entre os vértices A e C e entre os vértices A e B , caracterizando-o como um multigrafo.

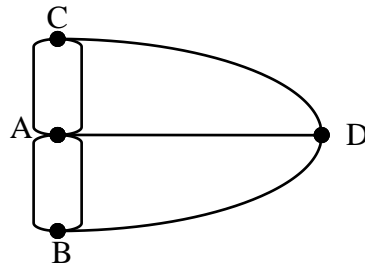


Figura 2.1.18: Multigrafo

2.2. Grafos Orientados Acíclicos (DAGs)

2.2.1. Definição

Um grafo é direcionado acíclico, se não tem circuitos. Isto é, um grafo orientado que não tem caminho fechado. A nomenclatura de grafos varia bastante conforme o autor, daí o nome “acíclico” (para “acyclic” em inglês) referindo-se a grafos sem circuitos.

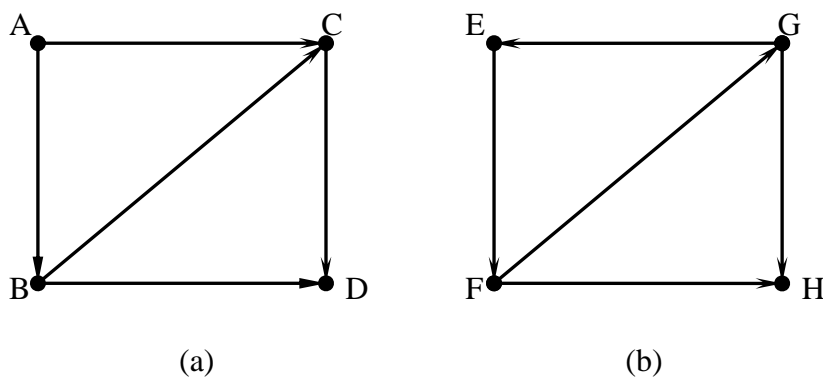


Figura 2.2.1: grafo direcionado acíclico (a) e direcionado não acíclico (b)

O primeiro grafo (2.2.1a) é DAG, enquanto que o segundo (2.2.1b) não é DAG, porque existe um caminho que forma um circuito ($E=\{E, F, G, E\}$).

Como foi definido no capítulo anterior, se $v \rightarrow w$, então “ v ” é antecessor ou pai de “ w ” e “ w ” é sucessor ou filho de “ v ”. O conjunto de pais de “ w ” é denotado por $pa(w)$ e o conjunto de filhos $ch(w)$.

Portanto, se existe um caminho orientado de “ v ” a “ w ”, então se pode dizer que “ v ” é algum ancestral de “ w ” e “ w ” é chamado de descendente de “ v ”. O conjunto de ancestrais de “ w ” é denominado $an(w)$ e o conjunto de descendentes como $de(w)$.

Nota: $an(v)$ coincide com a definição de $fti(v)$ e $de(v)$ coincide com a definição de $ftd(v)$.

Essas definições (pai, filho, ancestrais e descendentes) podem facilmente serem estendidas para aplicação ao conjunto de vértices. Por exemplo, para o conjunto S igual ou contido em V nós definimos $pa(S) = \{\cup_{v \in S} pa(v)\} \setminus S$, Isto é, como um conjunto de vértices que não estão em “ S ” e que são pais de um vértice em “ S ”.

Não é difícil mostrar que a ausência de algum ciclo orientado é equivalente a existência de alguma ordenação dos vértices $\{v_1, \dots, v_n\}$, tal que $v_i \rightarrow v_j$, implica $i < j$. Em outras palavras, existem vértices numerados de modo que as setas apenas partem do vértice com menor valor para o de maior numeração. Naturalmente, a numeração não é necessariamente única.

Os vértices podem representar variáveis e podem ser rotuladas v_1, \dots, v_n tal que v_i está antes de v_{i+1} para $i = 1, \dots, n - 1$.

2.2.2. Grafos de Cadeia (Chain Graphs)

Grafo de cadeia é um grafo composto por subgrafos não orientados, denominados “blocos” que podem (ou não) serem conectados por arcos, de forma que nenhum destes arcos participe de um circuito.

Portanto, Grafos de cadeia são grafos compostos por vários blocos, em que cada bloco (B_1, \dots, B_r) contem apenas arestas não orientadas. Se dois blocos são conectados, eles são conectados por uma seta (arco). Além disso, todas as setas entre dois blocos quaisquer devem ter a mesma direção (caso contrário, seria fácil a formação de um ciclo orientado).

Pode-se construir um grafo orientado do qual os vértices são os componentes de G e onde dois vértices são conectados por uma seta sempre que os dois componentes são conectados em G . Portanto, o grafo formado por esses componentes é um DAG.

Este resultado pode ser explorado de várias maneiras. Por exemplo, os blocos podem ser enumerados e ordenados, de modo que os arcos são sempre do bloco com número inferior ao de maior número. De novo, a salientar que a enumeração não é necessariamente única.

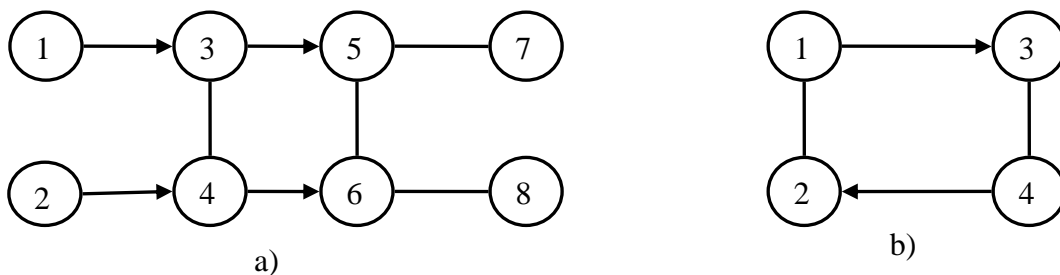


Figura 2.2.2: grafo de cadeia (a) e grafo não de cadeia (b)

Na figura acima (2.2.2) temos em (a) um grafo de cadeia com os blocos: $B_1 = \{1\}$, $B_2 = \{2\}$, $B_3 = \{3,4\}$ e $B_4 = \{5,6,7,8\}$, em (b) o grafo forma um ciclo, portanto, não é grafo de cadeia.

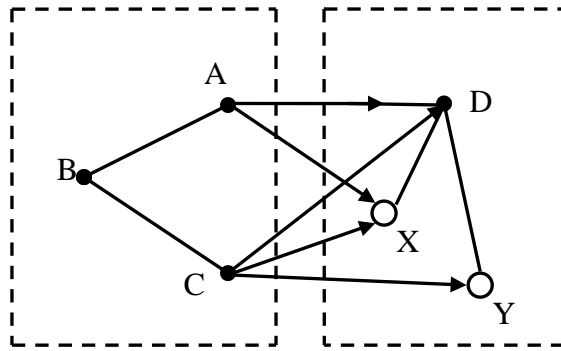


Figura 2.2.2c: grafo de cadeia.

Os dois blocos, $\{A, B, C\}$ e $\{D, X, Y\}$, são mostrados em caixas. Arestas entre as caixas são orientadas (arcos) e arestas dentro das caixas são não orientadas.

Por exemplo, na figura 2.2.2d temos um ciclo $\{X, Y, Z, X\}$, mas ele é um ciclo não orientado. É fácil ver que todos os ciclos são iguais a esse. Mas se houvesse uma linha ao invés de uma seta (arco) entre K e Z, a componente $\{K, X, Z, K\}$ seria um ciclo orientado, e assim o grafo em 2.2.2d não seria um “grafo de cadeia”.

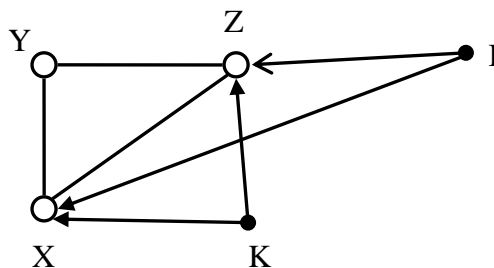


Figura 2.2.2d: Outro grafo de cadeia.

A classe de “grafos de cadeia” inclui grafos não orientados e DAGs como casos especiais.

Outro conceito que se utiliza com muita frequência em grafos de cadeia é o “componente”, que não é nada mais a que dá origem a uma (possivelmente) partição fina das variáveis (vértices) do que a que foi dada pelos blocos. Isto é, cada bloco é uma união de um ou mais componentes.

Por exemplo, os componentes da Figura 2.2.2d são $\{I\}$, $\{K\}$ e $\{X, Y, Z\}$. Portanto, este grafo pode ter sido gerado a partir varias estruturas de blocos:

1. $B_1 = \{I, K\}$ e $B_2 = \{X, Y, Z\}$,
2. $B_1 = \{I\}$, $B_2 = \{k\}$, e $B_3 = \{X, Y, Z\}$, ou
3. $B_1 = \{k\}$, $B_2 = \{I\}$, e $B_3 = \{X, Y, Z\}$

Igualmente, para o grafo da figura 2.2.2a temos as seguintes estruturas de bloco possíveis:

1. $B_1 = \{1, 2\}$, $B_2 = \{3, 4\}$ e $B_3 = \{5, 6, 7, 8\}$,
2. $B_1 = \{1\}$, $B_2 = \{2\}$, $B_3 = \{3, 4\}$ e $B_4 = \{5, 6, 7, 8\}$.

Uma vez que os componentes podem ser identificados a partir do grafo, segue-se que não é necessário desenhar as caixas do grafo (ver figura 2.2.2e). Toda propriedade matemática e estatística pode ser derivada apenas a partir do grafo, sem especificar a estrutura do bloco.

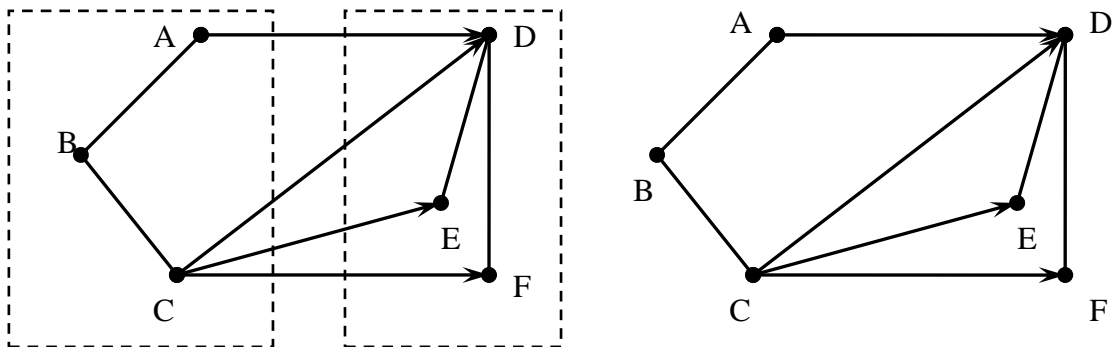


Figura 2.2.2e: Desencadeando um grafo em blocos

O grafo a esquerda mostra um grafo de cadeia com os blocos desenhados, e o da direita é o mesmo sem os blocos.

2.2.3. Grafo Moralizado (Moral Graph)

Definimos o grafo moralizado G^m como sendo a versão não orientada do grafo de cadeia G acrescido de arestas não direcionadas entre todo o par de vértices que possua filho em comum e que não sejam adjacentes.

Dado um DAG $G = (V, A)$, do qual construímos o seu grafo moral (G^m) não orientado, casando (unindo) os seus pais e apagando as orientações, ou seja,

- i. Para cada $v \in V$, conectar todos $pa(v)$ com linhas (aresta).
- ii. Substituímos todas as setas ou arcos em A com linhas.

Portanto, chamamos ao grafo resultante (G^m) de “moral graph” correspondente ao G (figura 3.3b).

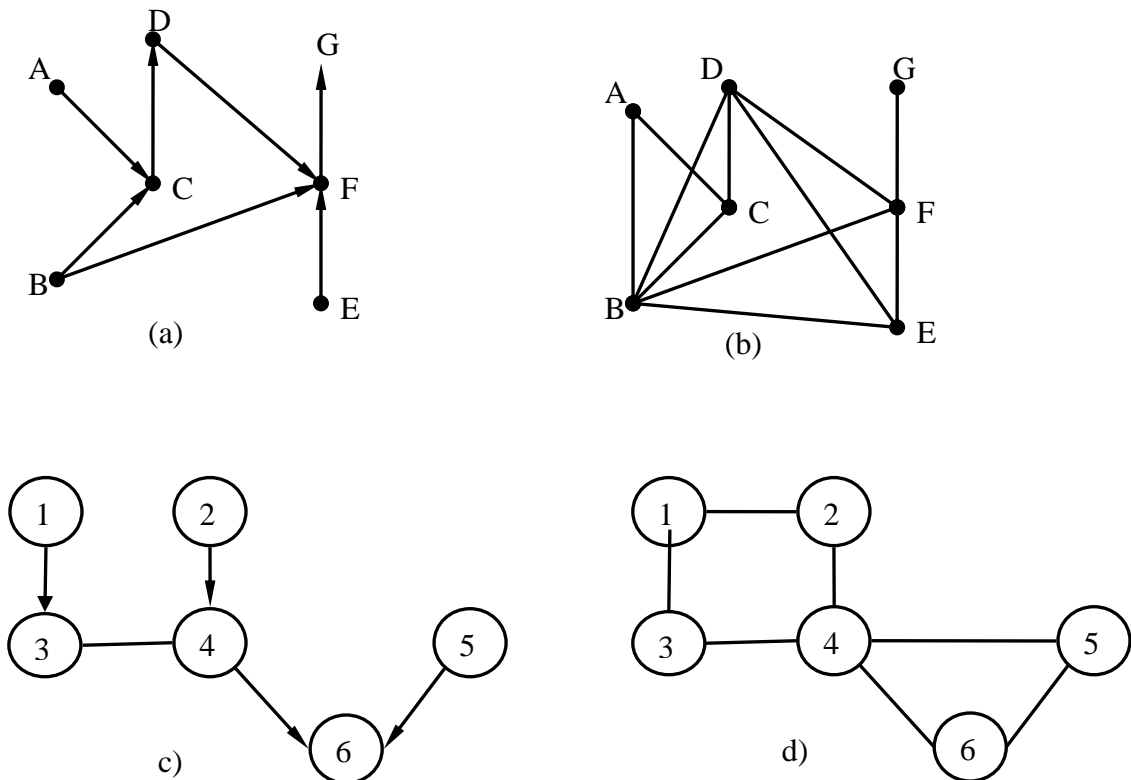


Figura 2.2.3: grafo orientado acíclico (a, c) e o seu grafo moralizado (b, d)

Na figura acima, o grafo em (a) mostra um DAG “G” e em (b) mostra o seu grafo moralizado G^m , por outro lado, em “c” temos cadeia grafo e seu grafo moralizado em “d”. Ambos G^m são formados pelo casamento dos pais em G e em seguida a eliminação das orientações nas arestas. Em G da figura 2.2.3a vem-se que $pa(C) = \{A, B\}$ e $pa(F) = \{B, D, E\}$.

Note que (em c) o bloco $\{3, 4\}$ é considerado filho de $\{1\}$ e $\{2\}$ o que justifica a ligação no grafo moralizado desses dois vértices (1 e 2).

Capítulo 3

Conceitos Probabilísticos

Modelos grafos constituem uma ferramenta natural para lidar com problemas que envolvem incerteza e complexidade. Podem, resumidamente, ser caracterizados como uma fusão perfeita entre a teoria de probabilidades e a teoria de grafos. Portanto, modelos grafos são meros grafos nos quais, os nós são variáveis aleatórias e a ausência de arcos representa alguma hipótese de independência condicional.

Vale recordar, que existem duas classes de modelos gráficos: os baseados em grafos não direcionados e aqueles baseados em grafos direcionados. Os primeiros incluem os modelos de campos aleatórios Markovianos e os últimos, as denominadas redes Bayesianas e os diagramas de influência.

3.1. Independência Condicional em Grafos

Grafos são instrumentos adequados para representar as relações de dependência condicional em variáveis. Seja $X = (X_1, X_2, \dots, X_k)$ um vetor de variáveis aleatórias e independentes.

Usaremos a notação $x_i \perp\!\!\!\perp x_j / S$ para referir que x_i é condicionalmente independente de x_j dado o conjunto S , e $x_i \perp\!\!\!\perp x_j / K \setminus \{i, j\}$ para x_i condicionalmente independente de x_j dado o resto $(K \setminus \{i, j\})$.

Um grafo que represente as relações de dependência condicional de X é um grafo não orientado $G=(V,A)$ no qual $V = \{1,2, \dots, k\}$ e (i, j) não está no conjunto de arestas “A” se, e somente se, $x_i \perp\!\!\!\perp x_j / x_k \setminus \{i, j\}$.

Exemplo: tome $V = 4$, e considere a função densidade de $X = (X_1, X_2, X_3, X_4)$; como sendo $f_X(x) = \exp(u + x_1 + x_1x_2 + x_2x_3x_4)$, $x = (x_1, x_2, x_3, x_4)$, $0 < x_i < 1$, $i = 1, 2, 3, 4$, em que a constante u garante a densidade integrada para 1. A aplicação direta do critério da fatoração implica que $f_X(x)$ pode ser fatorada como

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

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

- i. $x_1 \perp\!\!\!\perp (x_3, x_4) / x_2$
- ii. $x_1 \perp\!\!\!\perp x_4 / (x_2, x_3)$ e $x_1 \perp\!\!\!\perp x_3 / (x_2, x_4)$

E conseqüentemente o grafo correspondente é dado pela figura 3.1.

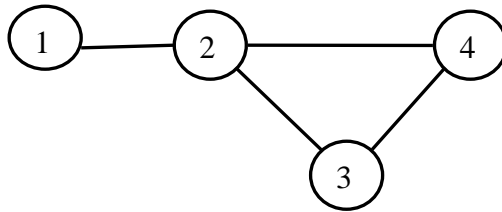


Figura 3.1: grafo ilustrando a relação de dependência de quatro variáveis

Note que o critério de fatoração de acordo com Whittaker (1990) diz que: as variáveis aleatórias Y e Z são condicionalmente independentes dado X ($Y \perp\!\!\!\perp Z / X$), se, e somente se, existe uma função g e h tal que $f_{XYZ}(x, y, z) = g(x, y)h(x, z)$ para todo y e z e todo x com $f_x(x) > 0$.

3.1.1. Propriedades de Markov

Um Processo de Markov é um processo estocástico, no qual as distribuições de probabilidade para o seu desenvolvimento futuro dependem somente do estado presente, portanto, não levando em consideração como o processo chegou a tal estado.

Uma cadeia de Markov é um processo estocástico que apresenta a propriedade de Markov, chamada assim, em homenagem ao matemático Andrei Andreyevich Markov. Note que uma cadeia de Markov é uma seqüência $v_1, v_2, v_3, \dots, v_n$ de variáveis aleatórias.

A propriedade de Markov faz o uso da Teoria de Grafo ao considerar que os vértices são variáveis aleatórias envolvidas num problema e que os arcos representam uma relação casual entre as variáveis.

Se a distribuição de probabilidade condicional de v_{n+1} nos estados passados, é uma função apenas de v_n , então:

$$P(v_{n+1} = i \mid v_0, v_1, v_2, \dots, v_n) = P(v_{n+1} = i \mid v_n),$$

Em que v é alguma variável (estado do processo). A identidade acima define a *propriedade de Markov*.

A definição desta propriedade, também chamada de “memória markoviana”, tem como principal característica a suposição de falta de memória. Isto é quando se sabe o atual estado do processo (de um vértice), informações anteriores são irrelevantes nas inferências dos estados futuros.

Em outras palavras, os estados anteriores são considerados irrelevantes para a predição dos estados seguintes, desde que o estado atual seja conhecido.

A propriedade de Markov no grafo, tanto no caso de grafo não orientado como no caso grafos orientados acíclicos, tem sido assunto de muitas pesquisas recentes. Em muitos casos, ela é utilizada para representar independência dois-a-dois (pairwise).

Podem-se distinguir três propriedades de Markov, nomeadamente, propriedades de Markov para pares de variáveis (dois-a-dois), Propriedade de Markov local e propriedade de Markov global.

- Propriedade de Markov para pares de variáveis ou vértices (dois-a-dois)

Nesta propriedade, pares de variáveis (vértices) não adjacentes no grafo são condicionalmente independentes dada a terceira ou o resto das variáveis. Ou seja, $x_i \perp x_j / k \setminus \{i, j\}$.

Considerando a figura abaixo (figura 3.1.1a), temos que $x_1 \perp x_3 / (x_2, x_4, x_5, x_6)$ ou $x_4 \perp x_6 / (x_1, x_2, x_3, x_5)$.

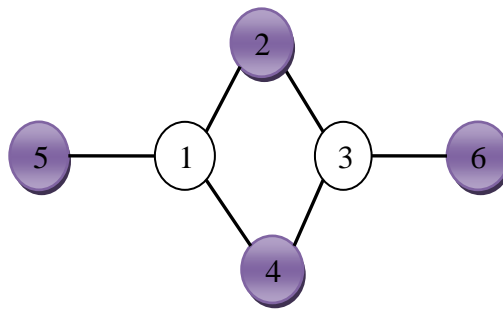


Figura 3.1.1a. Propriedade de Markov para pares de variáveis

- Propriedade de Markov Local:

Esta se caracteriza por condicionar somente sobre os vértices adjacentes, assim qualquer vértice é independente de todos os restantes. Ou seja, para cada vértice “i”, sendo “a” o conjunto de vértices adjacentes a “i” e sendo “b” o conjunto dos restantes, temos $x_i \perp x_b / x_a$ onde $b = k \setminus \{i\} \cup a$.

Considerando o grafo abaixo (figura 3.1.1b) temos

$$x_4 \perp x_b / x_a \text{ onde } a = (x_1, x_3) \text{ e } b = (x_5, x_2, x_6)$$

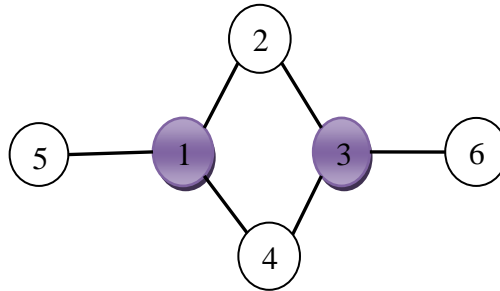


Figura 3.1.1b. Propriedade de Markov Local

- Propriedade de Markov Global:

Nesta, dois subconjuntos quaisquer de variáveis (vértices), 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 conjunto “a”, então $x_b \perp x_c / x_a$.

Considerando o grafo da figura 3.1.1c, temos

$$x_b \perp x_c / x_3 \text{ onde } c = (x_6) \text{ e } b = (x_5, x_1, x_2, x_4)$$

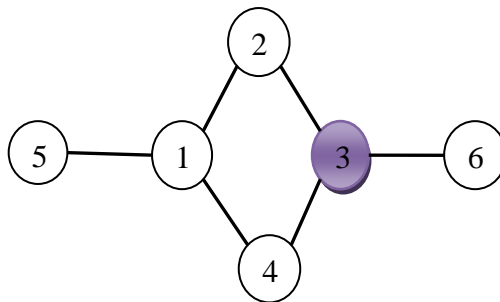


Figura 3.1.1c. Propriedade de Markov Global

3.1.2. Propriedades de Markov em Grafos Orientados Acíclicos (DAGs)

Uma pergunta que ficou muito tempo sem resposta era se existia uma propriedade de Markov global equivalente para DAGs (grafos orientados acíclico).

A propriedade de Markov dois-a-dois (pairwise), afirma que $B \perp\!\!\!\perp D \mid \{A, C\}$; mas isso também assegura que $B \perp\!\!\!\perp D \mid C$? Intuitivamente, isto parece provável.

Por exemplo, na figura 3.1.2a, não existe seta (ligação ou arco) de B para D.



Figura 3.1.2a: grafo orientado acíclico (DAG) simples

Sabe-se que todas as independências condicionais podem ser encontradas a partir das variáveis (vértices) que, quando fixadas, interrompem o fluxo de dependência.

Estas independências condicionais são propriedades dos DAGs conhecidas por d-separação, onde d se refere à *direction-dependent*.

A d-separação tem a característica de bloquear a comunicação entre dois conjuntos disjuntos a partir de um terceiro subconjunto também disjunto destes. O termo bloquear é referente a interromper o fluxo de dependências.

De acordo com Edwards (1995), atualmente existem dois critérios de formulações diferentes sobre propriedade de Markov para DAGs. A formulação original é devido a Pearl (1986a, 1986b) e Verma e Pearl (1990a, 1990b); pouco tempo depois, Lauritzen *et al.* (1990) apresentaram uma formulação alternativa, ambos citados por Edwards (1995).

Em primeiro lugar, vamos aqui analisar o critério da versão mais recente, que usa grafos moralizados. Como anteriormente no capítulo dois, a partir de um DAG $G = (V, A)$, construímos o seu grafo moral (G^m) não orientado, casando (unindo) os seus pais e apagando as orientações (ver figuras 3.1.2b e 3.1.2c abaixo)

Agora vamos supor que queremos verificar se $v_i \perp v_j / S$ (para um conjunto S contido em V).

A propriedade global de Markov para grafos orientados se aplica no sentido de que se S separa v_i e v_j no grafo moralizado, então $v_i \perp v_j / S$.

- Assim, se S separa v_i e v_j em G^m_A , onde A é o conjunto ancestral de v_i e v_j união com o conjunto S , então, $v_i \perp v_j / S$.

Suponha que queremos verificar se $C \perp F / D$ sob o grafo da figura 3.1.2b. Para isso, primeiro, a partir da figura 3.1.2b devemos construir o seu grafo moralizado (figura 3.1.2c).

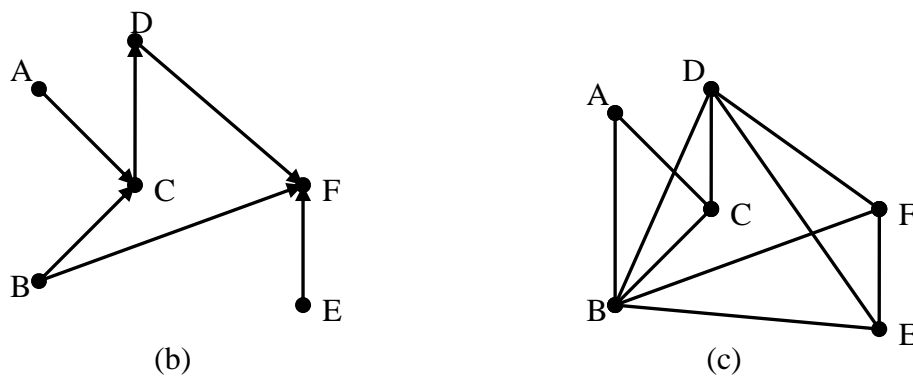


Figura 3.1.2b, c: grafo orientado acíclico (b) e o seu grafo moralizado (c)

Aplicando o critério de d-separação em figura 3.1.2 (c), D não separa C de F , assim, não é verdade dizer que $C \perp F / D$.

Esse critério pode ser facilmente estendido para um conjunto maior de variáveis (vértices).

Em resumo, a versão orientada da propriedade global de Markov afirma que, para três conjuntos disjuntos S_1 , S_2 e S_3 , vale $S_1 \perp\!\!\!\perp S_2 / S_3$, sempre que S_3 separa S_1 e S_2 em G^m_A ; onde $A = \text{an}^+(S_1 \cup S_2 \cup S_3)$.

A versão original do critério (propriedade de Markov em DAGs) focaliza os diferentes caminhos entre vértices. Por exemplo, em grafos não orientados:

- A existência de caminhos entre v e w , por exemplo, indica que eles são marginalmente dependentes;
- Se estivermos interessados em dependência condicional de v e w dado o conjunto S , e se o caminho não contém um vértice em S , isso (ainda) indica dependência condicional;
- Se o caminho contiver o tal vértice, então não está claro se aplica ou não a independência condicional. Só estará claro se todos os caminhos entre v e w contiverem um nó em S . assim S vai separar v e w , de modo a que estes sejam condicionalmente independente dado S .

Um argumento similar aplica-se para DAGs, mas aqui, antes, é crucial distinguir o tipo de configuração do caminho.

Num grafo direcionado, chama-se “*collider*” a um vértice que tiver duas ou mais setas convergentes, caso contrario o vértice é “*noncollider*”.

Considere o DAGs mostrado na figura 3.1.2d e (e), ambos têm caminhos de V para W .

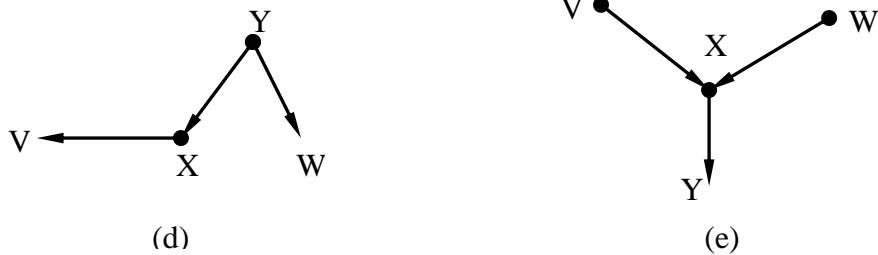


Figura 3.1.2d, e: Dois grafos orientados acíclicos ilustrando o caminho entre V e W

Na figura acima, tenho que em (d) não há vértice “colliders”, em (e) existe um vértice *collider* (X).

Agora, examinemos varias relações de (in)dependência entre V e W que estão associados com estes caminhos, mantendo em mente que, quando estes grafos são incorporados em grafos maiores, a independência encontrada aqui podem desaparecer, mas a dependência continua assegurada.

Na figura 3.1.2d, o caminho não contém “colliders” e tem-se que V não é independente de W, mas que $V \perp W / X$ e $V \perp W / Y$.

- Portanto, podemos dizer que a cadeia de V a W nesta figura (3.1.2d) indica que V e W são marginalmente dependentes, mas que o caminho pode ser condicionalmente bloqueado pelos não “colliders” X ou Y.

Na figura 3.1.2e o caminho contém um “collider”, e temos que $V \perp W$, mas que V e W, dado X ou dado Y, não são independentes.

- Deste modo, para a figura (3.1.2e) a cadeia de V a W não indica dependência marginal, uma vez que contém um “collider”, no entanto, se condicionado sobre o “collider”, ou sobre os descendentes do “collider”, o caminho indica dependência entre V e W.

Juntando estas idéias, pode-se dizer que o caminho entre V e W pode ser ativado ou bloqueado. Estando ativo, significa ou indica uma dependência entre V e W.

Desta maneira, podemos concluir que uma cadeia é bloqueada se ou:

- i. tem vértice não “*collider*” que é condicionado, ou
- ii. ela tem um “*collider*” ligado que não é condicionado (e nenhum de seus descendentes é condicionado em ambos).

Voltando ao critério *d-separação* na sua formulação original (Pearl, 1986^a, 1986b, Verma e Pearl, 1990^a, 1990b, citado por Edwards, 1995) e considerando as cadeias entre os vértices em S_1 e S_2 , para definir o *d-separação* dos conjuntos S_1 e S_2 pelo S_3 , dizemos que S_3 bloqueia tal cadeia se ou

- (i) a cadeia tem um “*noncollider*”, digamos “ x ”, tal que $x \in S_3$, ou
- (ii) a cadeia tem um “*collider*”, digamos y , tal que $y \notin S_3$ e $de(x) \cap S_3 = \emptyset$.

Assim, este critério afirma que S_3 é *d-separação* de S_1 e S_2 se ele bloqueia todos os caminhos entre S_1 e S_2 .

3.1.3. Propriedade de Markov em Grafos de Cadeia

Embora os problemas com ordenações causais completas pareçam ser bastante raros em aplicações, ordenações parciais são frequentemente disponíveis. Por exemplo, um estudo epidemiológico poderá envolver as seguintes características de uma amostra de indivíduos:

1. Características familiares, como o genótipo parental.
2. Características genéticas, como o genótipo individual.
3. Características demográficas, como o sexo e grupo étnico.
4. Fatores sociais e econômicos, como a ocupação, estado socioeconômico, e experiência educacional.
5. Características de modo de vida, como o uso de tabaco, dieta, e exercícios físicos.
6. Características biológicas, como colesterol elevado e índice de massa corporal.

Claramente, as características familiares são antecedentes a características demográficas, as quais são antecedentes ao estilo de vida e características biológicas. Nisto pode ser razoável supor que os fatores socioeconômicos são antecedentes para o estilo de vida e características biológicas, mas esta é claramente uma hipótese não trivial.

Por exemplo, em estudos de ensaios clínicos se utiliza freqüentemente desenhos de semelhanças longitudinais: após a alocação ao tratamento aleatorizado, pacientes são seguidos ao longo de alguns intervalos adequados, e as informações são coletadas em pontos de tempo predeterminado. Para tais estudos, a ordenação das medições, dá origem a uma ordenação entre os blocos de variáveis, mas não a uma ordenação completa entre todas as variáveis.

Para captar este tipo de informação a priori, alguns trabalhos têm combinado os grafos não orientados e DAGs em um quadro único, o chamado bloco-recursivo ou grafo de cadeia (chain graph).

Esses grafos são baseados naquilo que é conhecido como uma dependência da cadeia, ou seja, se supõe que um conhecimento dá um compartilhamento das variáveis em uma lista ordenada do bloco, digamos $V = B_1 \cup B_2 \dots B_k$

Variáveis no mesmo bloco são supostamente concorrentes, isto é, a sua estrutura de associação é considerada simétrica, sem ordenação, como vimos no capítulo três. A figura 2.2.2a é algum exemplo de grafo semelhante.

Correspondente à estrutura de bloco assumimos que a densidade conjunta $f(V_1, \dots, V_2)$ é fatorizada na forma:

$$f(B_1)f(B_2|B_1) \dots f(B_k|B_{k-1})$$

Se estiver faltando uma linha entre dois vértices v e w no mesmo bloco B_i , ou um arco está faltando em $v \in B_j$ para $w \in B_i$, onde $j < i$, então isso significa que

$$v \perp\!\!\!\perp w / B_1 \cup B_2 \dots \cup B_i$$

Sendo esta uma versão de Propriedade de Markov *Pairwise* (dois-a-dois) para grafos de cadeia (*Chain graphs*), como explicado anteriormente.

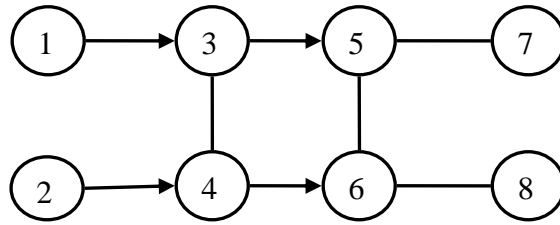


Figura 3.1.3a: grafos de cadeia

A partir da figura 3.1.3a, podemos construir o seguinte DAG dos seus componentes:

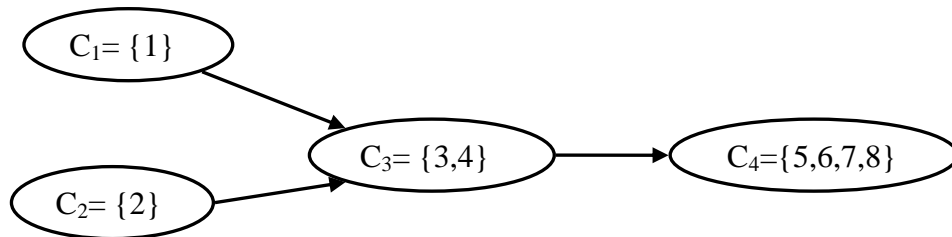


Figura 3.1.3b: DAG dos componentes da cadeia em “a”

Da figura acima (3.1.3b), temos que o pai de C_4 é C_3 e os seus não descendentes são $C_1 \cup C_2 \cup C_3$.

Por outro lado usando a propriedade de independência condicional de Markov verificamos, por exemplo, que:

$$\triangleright C_4 \perp\!\!\!\perp \{C_1, C_2\} / C_3.$$

E pela figura 3.1.3a, temos que,

$$\triangleright \{5, 7\} \perp\!\!\!\perp \{8\} / \{3, 4, 6\}$$

$$\triangleright \{5, 7\} \perp\!\!\!\perp \{4\} \mid \{3, 6\} \dots$$

Aplicando a fatoração da função de densidade

$$f(x_1, \dots, x_8) = f(x_1)f(x_2)f(x_3, x_4|x_1, x_2)f(x_5, x_6, x_7, x_8|x_3, x_4)$$

A estrutura do bloco nos dá informações sobre modelos alternativos, em especial sobre a falta de arestas; por exemplo, na figura 3.1.3c não se sabe se, o gênero e raça forem conectados, seria por uma linha ou uma seta. O conhecimento da estrutura do bloco é necessário para casos como esses.

A estrutura de bloco codifica informações prévias sobre o sistema, enquanto que os componentes são modelo específico.

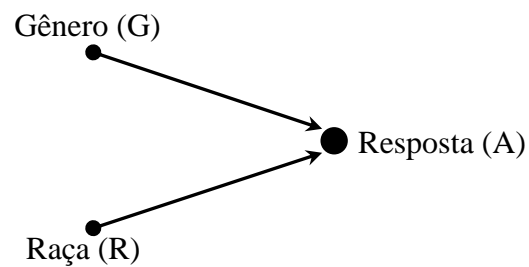


Figura 3.1.3c: Um grafo orientado mostrando que G é independente de R

Capítulo 4

Redes Bayesianas

4.1. Causalidade

A causalidade é uma condição segundo a qual uma causa produz efeito. Quando se tenta avaliar o impacto de novo acessório de um determinado produto no mercado, o rendimento de um time de futebol com a entrada de um novo atleta ou no que pode resultar a falta de cuidados com a saúde, o interesse recai sobre os efeitos que tais eventos podem causar.

Porém, se o problema é explicar por que as vendas caíram, por que o time não vence ou mesmo diagnosticar uma doença, a perspectiva volta-se as causas das ocorrências destes eventos.

Sob esta ótica, o grande desafio foi encontrar uma maneira de traduzir matematicamente a causalidade existente entre os eventos que compõem um dado problema de uma maneira realista, isto é, sem desprezar as incertezas inerentes a ele. Assim, o cálculo das probabilidades foi adotado.

Quando se raciocina com incerteza é importante saber se a informação sobre algum evento influencia a crença em outros, devido à criação e eliminação de relacionamentos de dependências.

As formas de propagação de evidência entre variáveis num DAG (rede causal) podem ter três tipos de conexões (Figura 4.1).

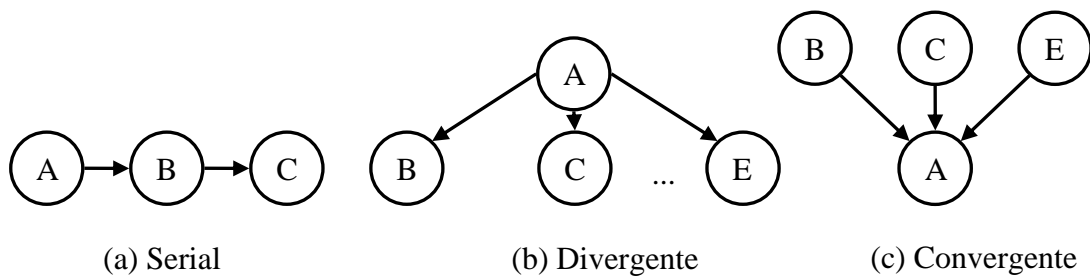


Figura 4.1: Tipos de Conexões numa Rede Causal

Na *conexão serial* (figura 4.1a), uma evidência em “A” influencia a crença em “B” que influencia a crença em “C”. De forma similar uma evidência em “C” se propaga para “A”. Em ambos os casos não há propagação de influência se “B” está instanciado (ou se o estado de “B” é conhecido) porque o canal entre “A” e “C” fica bloqueado, tornando-os condicionalmente independentes.

Portanto uma evidência pode ser transmitida através de uma conexão serial, exceto se o canal for bloqueado pelo conhecimento do estado de uma variável intermediária.

- *Câncer metastático* pode produzir *tumor no cérebro*, que é potencial causador de *forte dor de cabeça*. Confirmado o tumor, saber a ocorrência de *forte dor de cabeça* não altera a crença na existência ou não de *metástases* e vice-versa.

Na *conexão divergente* ((figura 4.1b) uma evidência em um ascendente de “A” influencia a crença sobre os filhos de A, exceto se A é instanciado. Assim uma evidência em “A” bloqueia o canal de comunicação com os seus filhos, tornando-os condicionalmente independentes. Diz-se que B, C,..., E são condicionalmente independentes, dado “A”.

A evidência pode ser transmitida através de uma conexão divergente a menos que a conexão seja instanciada.

Na *conexão convergente* (figura 4.1c) evidência em A ou em um dos seus descendentes influencia a crença nos pais de A, tornando-os condicionalmente dependentes devido ao princípio de *explaining away*. Se nada é conhecido sobre A,

exceto o que pode ser inferido a partir do conhecimento de seus pais B, C,..., E então eles são independentes, pois evidência em um deles não influencia a crença sobre os outros.

Portanto evidência só pode ser transmitida em uma conexão convergente se nem a variável na conexão ou qualquer dos seus descendentes recebeu alguma evidência. No exemplo em que *chuva* e *esguicho ligado* são causas potenciais de *pavimento molhado*, o qual é causa potencial de *queda acidental*, saber que houve uma queda torna *Chuva* e *esguicho ligado* dependentes entre si devido ao mecanismo de “*explaining away*”.

Segundo Jensen (2001) duas variáveis A e C numa rede casual são condicionalmente independentes se para todos os caminhos entre A e C existe uma variável intermediária B tal que:

- a) A conexão é serial ou divergente e o estado de B é conhecido (B instanciado ou o seu estado é conhecido);
- b) A conexão é convergente e nem B ou seus descendentes receberam qualquer evidência.

Para caracterização de situações de incerteza podemos utilizar grafos representando relações causais entre eventos. Como exemplo, considere o seguinte domínio:

Pela manhã meu carro não irá funcionar. Eu posso ouvir a ignição, mas nada acontece”. Podem existir várias razões para o problema.

- O rádio funciona então a bateria está boa.
- A causa mais provável é que a gasolina tenha sido roubada durante a noite ou que a mangueira esteja entupida.
- Também pode ser que seja o carburador sujo, um vazamento na ignição ou algo mais sério.

Para descobrir primeiro eu verifico o medidor de gasolina. Ele “indica $\frac{1}{2}$ tanque, então eu decido limpar a mangueira da gasolina”.

Este problema pode ser representado, de forma simplificada, pelos seguintes eventos: {sim, não} Gasolina? {sim, não} Mangueira limpa? {cheio, $\frac{1}{2}$, vazio} Medidor, {sim, não} Funcionando. Em outras palavras, os eventos são agrupados em variáveis que podem assumir alguns estados.

Nós sabemos que o estado Gasolina e o estado Mangueira entupida, possuem impacto direto no estado Funcionando, assim como Gasolina, tem impacto em Medidor.

Estas informações podem ser representadas como apresenta a figura 4.1d. Nesta figura também estão representadas as direções do impacto da variável. Neste caso, a maior certeza em uma causa, é movida na direção positiva.

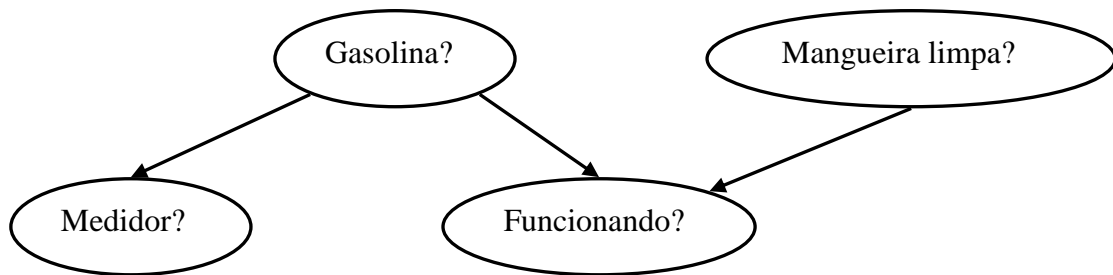


Figura 4.1d: Grafo representando uma rede de causalidade

O grafo da figura (4.1d) possibilita a realização de algumas conclusões. Obviamente, se eu sei que a mangueira está entupida (probabilidade 1), então a certeza do não funcionamento do carro aumenta.

Entretanto, suponhamos que minha situação é oposta. Isto é, eu sei que o carro não funciona e desejo encontrar o motivo.

A suspeita de que a gasolina pode ter sido roubada aumenta a expectativa de que o medidor esteja indicando vazio.

Lendo a informação do Medidor = $\frac{1}{2}$, diminui as expectativas de que o estado Gasolina seja a fonte do problema, guiando a aresta Gasolina/Funcionamento na direção

negativa. Uma vez que a expectativa do problema ter sido causado pela falta de gasolina diminui, cresce a expectativa sobre o estado Mangueira limpa. Neste instante é possível se concluir o seguinte:

“O problema não parece ser a Gasolina, então muito provavelmente deve ser a Mangueira”.

Para fazer estas escolhas, um agente deve inicialmente possuir preferências entre possíveis efeitos das ações a serem tomadas.

4.2. Redes Bayesianas

Uma rede Bayesiana é um par (G, Θ) , em que G é um grafo orientado Acíclico (DAG) e “ Θ ” é um conjunto particular de parâmetros. Este conjunto de parâmetros especifica as distribuições de probabilidade condicional associadas às variáveis representadas em “ G ”.

Por outro lado, Redes bayesianas constituem uma forma natural para representação de informações condicionalmente independentes.

Matematicamente, uma Rede Bayesiana é uma representação compacta de uma tabela de conjunção de probabilidades do universo do problema. Do ponto de vista de um especialista, Redes Bayesianas constituem um modelo gráfico que representa de forma simples as relações de causalidade entre as variáveis de um sistema.

Uma rede bayesiana consiste no seguinte:

- Um conjunto de variáveis e um conjunto de arestas orientadas entre as variáveis.
- Cada variável tem um conjunto finito de estados mutuamente exclusivos.
- As variáveis unidas por arestas orientadas formam um DAG.
- Para cada variável “ A ” com parentes B_1, \dots, B_n , existe anexada uma tabela potencial $P(A|B_1, \dots, B_n)$.

Note que, caso “A” não possua um pai, a tabela de probabilidades é reduzida para uma probabilidade incondicional ou a priori $P(A)$. Uma vez definida a topologia da rede, basta especificar as probabilidades dos nós (vértices) que participam em dependências diretas, e utilizar estas para computar as demais probabilidades que se deseje.

Portanto, para o DAG na figura 4.2a, as probabilidades a priori de $P(A)$ e $P(B)$ devem ser especificada.

A definição de redes Bayesianas, não se refere apenas à casualidade, e não há exigência de que sempre as ligações representem impacto causal. Além disso, solicita-se que a propriedade d-separação seja implícita pelo suporte da estrutura.

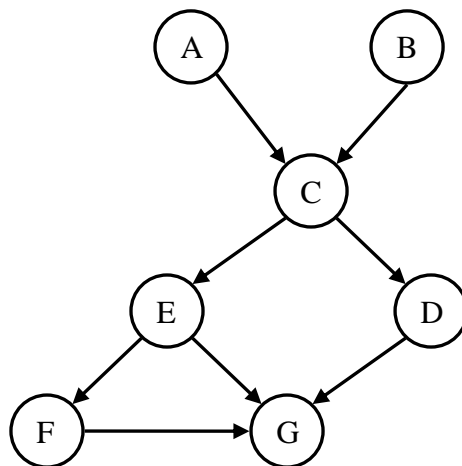


Figura 4.2a: grafo orientado acíclico (DAG)

As probabilidades para especificar são: $P(A)$, $P(B)$, $P(C|A,B)$, $P(E|C)$, $P(D|C)$, $P(F|E)$ e $P(G|D,E,F)$.

$P(C|A)$ e $P(C|B)$ sozinhos não nos dizem nada a respeito do impacto do conjunto $\{A, B\}$ sobre C .

Portanto, esses impactos podem se reforçar ou anular. Assim, necessitamos de $P(B|A,C)$.

Isto também significa que, se "A" e "B" são d-separados, dada a evidência "e", então cálculo de probabilidade utilizado para rede Bayesiana deve dar $P(A|B, e) = P(A|e)$.

A Inferência Bayesiana é relativamente simples quando envolve apenas duas variáveis.

Porém, quando a quantidade de variáveis se eleva, tal inferência torna-se muito complexa e às vezes sem valor prático. É neste momento em que as Redes Bayesianas se inserem ao problema da inferência Bayesiana, através da condição Markoviana atribuída às variáveis aleatórias envolvidas.

Na obtenção da estrutura da rede bayesiana, é preciso verificar se o relacionamento entre as variáveis está de acordo com o esperado, portanto, muitas das vezes é preciso a introdução de vértices auxiliares.

Por exemplo, existem dois testes para diagnosticar gravidez: Teste de sangue (*TS*) e Teste de urina (*TU*).

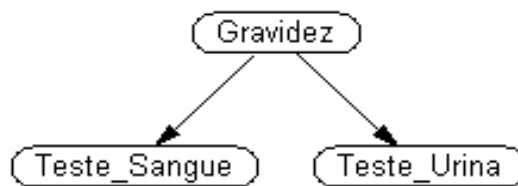


Figura 4.2b: Grafo ilustrando teste de gravidez

Se soubermos da gravidez, o modelo acima (figura 4.2b) diz que TS não nos informa nada sobre TU. Isso é falso, pois quanto maior TS, maior TU. Em outras palavras, o modelo deve refletir as relações de d-separação.

Na realidade, o que afeta os testes não é a gravidez, mas os hormônios. Assim, deve-se adicionar uma variável que reflete o nível hormonal (figura 4.2c).

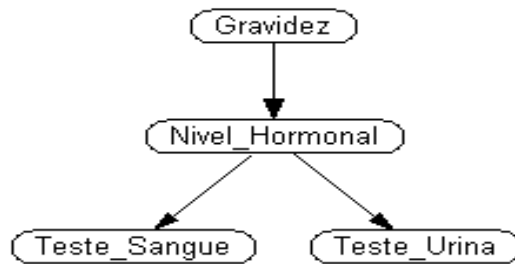


Figura 4.2c: Grafo ilustrando teste de gravidez com um vértice auxiliar

Por outro lado, se uma variável de uma rede bayesiana possui muitos pais (figura 4.2d), pode ser muito difícil obter sua tabela de probabilidades condicionais. Podemos então agrupar as causas em “departamentos”, este processo é conhecido por *Divorcing* (separação de pais).

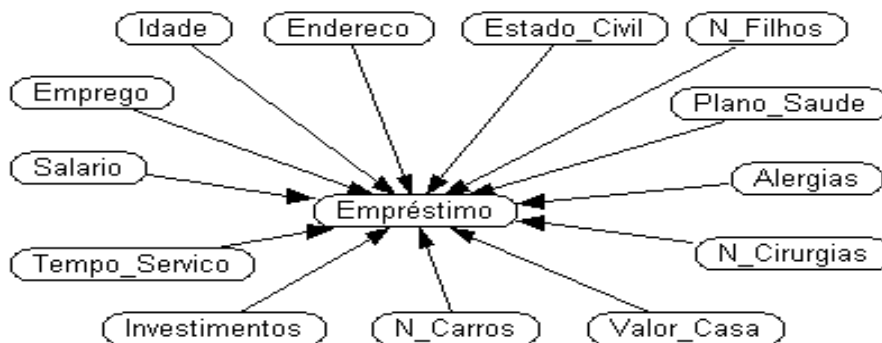


Figura 4.2d: Exemplo de rede bayesiana com muitos pais

Assim, tomando em conta a figura acima, para um banco decidir emprestar dinheiro a um cliente, ele deve observar dados pessoais (idade, endereço, estado civil e número de filhos), dados de saúde (número de cirurgias, alergias, plano de saúde), dados financeiros (número de carros, preço da casa, investimentos), dados trabalhistas (emprego, salário, status), etc...

Portanto, agrupando estas categorias temos o seguinte grafo:

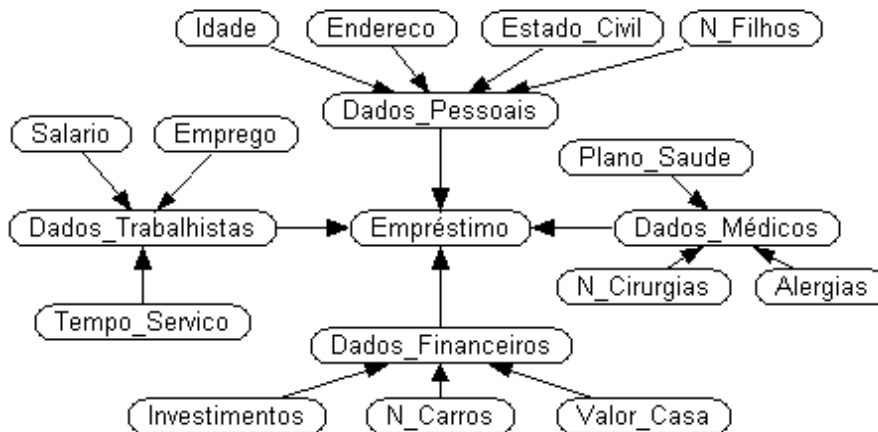


Figura 4.2e: rede bayesiana depois da separação de pais

4.2.1. Inferência em Redes Bayesianas

Considere um Exemplo de Rede Bayesiana abaixo em que nele ilustra a seguinte situação.

Exemplo 1: Imagine que você comprou um alarme para sua casa que pode ser disparado em duas situações: quando ocorre um terremoto (T) e quando um ladrão (L) tenta entrar na sua casa. Além disso, seus dois vizinhos John (J) e Mary (M) ligarão para o seu celular caso ouçam o alarme (A). Obviamente a situação não conta com o teor da conversa no telefone, mas sim com o fato de ser identificada uma ligação.

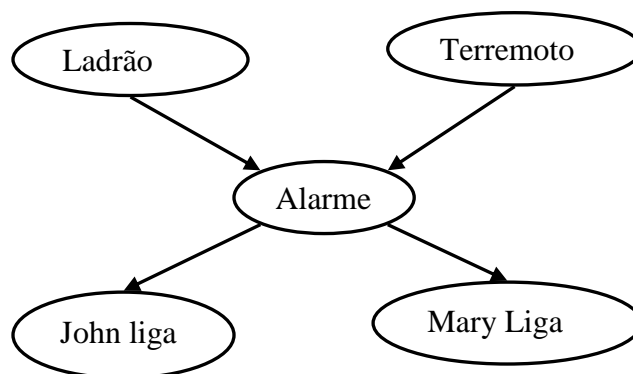


Figura 4.2.1: Exemplo de Rede Bayesiana

De acordo com a figura 4.2.1, Ladrão (L) e Terremoto (T) não têm probabilidade condicional, Alarme (A) é condicional a L e E, John liga (J) é condicional a “A” e por fim Mary liga (M) é condicional a “A”, as probabilidades para que esses eventos ocorram serão os seguinte:

$$P(B) = 0.001; P(E) = 0.002$$

Tabela 4.1: valores de probabilidade dos eventos (V = verdadeiro e F = Falso)

L	T	P(A)
V	V	0.95
V	F	0.94
F	V	0.29
F	F	0.001

A	P(J)
V	0.90
F	0.05

A	P(M)
V	0.70
F	0.01

A utilização de inferência em Redes Bayesianas se refere à tarefa de determinar a distribuição de probabilidade posterior de um grupo de variáveis X. Para isso, utilizamos um conjunto de variáveis de evidência “e”, ou seja, eventos observados, e um conjunto de variáveis não-evidenciadas Y, chamadas de variáveis escondidas.

Desta forma, uma inferência seria questionar qual a distribuição de probabilidade posterior $P(X | e)$, ou seja, o sistema calcula a probabilidade de uma variável de consulta (query variables), dada uma evidência forte (hard evidence). Isto é:

$$P(\text{Variável consulta} / \text{variáveis evidência})$$

Em outras palavras, inferência em redes Bayesianas é o procedimento realizado através de cálculos probabilísticos que utiliza as informações existentes na Rede Bayesiana, para a construção de uma distribuição de probabilidades posteriores para os objetos de interesse dadas as probabilidades a priori.

No exemplo dado, uma inferência a ser resolvida seria calcularmos qual a distribuição de probabilidade posterior de um ladrão estar tentando assaltar a sua casa,

dado que John e Mary fizeram uma ligação para você. Em outras palavras, realizarmos o seguinte cálculo:

$$P(\text{Ladrão} \mid \text{Johnligou} = V, \text{Maryligou} = V)$$

Na prática, é muito difícil calcular essa probabilidade de forma exata, ou seja, sem utilizar algoritmos numéricos de aproximação.

Todavia, inicialmente é importante termos em mente a seguinte equação, que representa a probabilidade de uma série de eventos ocorrerem, dados seus condicionais, ou seja, seus parentes no grafo da Rede Bayesiana:

$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{parentes}(x_i))$$

Além disso, para calcularmos $P(X \mid e)$ podemos utilizar a seguinte equação:

$$P(X \mid e) = \alpha P(X, e) = \alpha \sum_Y P(X, e, Y)$$

Nesta equação vemos que podemos transformar a probabilidade condicional ao evento “e” em um somatório de probabilidades que leva em conta todos os valores da variável escondida y, ou seja, $\alpha P(X, e, y) + \alpha P(X, e, \sim y)$, onde α é uma constante de normalização e ‘ $\sim y$ ’ é o complementar de y.

Voltando ao exemplo dado e considerando a equação acima citada, podemos calcular a probabilidade $P(L \mid J, M)$ da seguinte forma:

$$P(L \mid J, M) = \alpha P(L, J, M) = \alpha \sum_T \sum_A P(L, T, A, J, M)$$

Em que as variáveis escondidas, neste caso, são T e A, uma vez que tanto John quanto Mary dependem do Alarme e o Alarme, por sua vez, depende do Terremoto e do Ladrão, estando tanto o Alarme quando o Terremoto escondidos no problema.

Em outras palavras, não importa se houve terremoto ou não e não importa se o alarme tocou ou não. Simplesmente queremos saber se há um ladrão em casa, dado que John e Mary fizeram uma ligação.

Utilizando a equação $P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parentes}(x_i))$ na probabilidade final temos o seguinte:

$$P(L|J, M) = \alpha \sum_T \sum_A P(L)P(T)P(A|L, T)P(J|A)P(M|A)$$

Por fim, isolando as constantes que não fazem parte dos somatórios temos o seguinte:

$$P(L|J, M) = \alpha P(L) \sum_T P(T) \sum_A P(A|L, T)P(J|A)P(M|A)$$

Assim, chegamos a uma equação que mostra que a probabilidade posterior desejada pode ser calculada através de um produtório que tem como fatores somatórios em relação às variáveis escondidas. Para facilitar, podemos dizer que a equação pode ser expandida para o seguinte:

$$P(L | J, M) = \alpha (P(L) \cdot (P(T) \cdot (P(A | L, T) \cdot P(J | A) \cdot P(M | A) + P(\bar{A} | L, T) \cdot P(J | \bar{A}) \cdot P(M | \bar{A})) + P(\bar{T}) \cdot (P(A | L, \bar{T}) \cdot P(J | A) \cdot P(M | A) + P(\bar{A} | L, \bar{T}) \cdot P(J | \bar{A}) \cdot P(M | \bar{A}))))$$

Podemos também simbolizar esta expressão através duma árvore, que facilita a visualização, na qual as probabilidades de cada ramo podem ser determinadas através das tabelas de probabilidades mostradas na definição do problema.

É importante termos em mente que as variáveis que estamos tratando são binárias e, dessa forma, se não fosse binárias a expansão da somatória seria maior que esta. Além disso, estamos considerando poucas variáveis.

Realizando o cálculo teremos que $P(L | J, M) = \alpha 0,00059224$. Para normalizar o valor, fazemos o mesmo cálculo para $P(\bar{L} | J, M)$, que nos dará $P(\bar{L} | J, M) = \alpha 0,0014919$.

Assim, normalizando os valores temos:

$$P(L | J, M) = 0,00059224 / (0,00059224 + 0,0014919) = 0,284$$

$$P(\bar{L} | J, M) = 0,0014919 / (0,00059224 + 0,0014919) = 0,716.$$

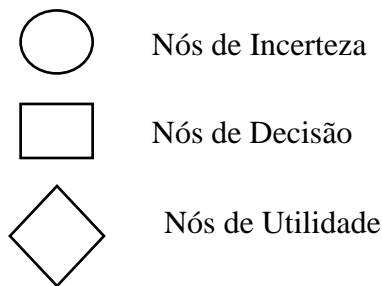
Concluimos, portanto, que há 28,4% de chance de haver um ladrão em casa, dado que John e Mary fizeram uma ligação.

4.3. Diagrama de Influência

Os diagramas de influência são redes bayesianas com nós especiais. Portanto, eles são considerados como caso especial de redes Bayesianas, que possuem nós de incerteza (chance nodes), nós decisão e nós de utilidade (nós de consequência).

- **Nós de incerteza** são variáveis aleatórias (ovais). Cada nó “A” tem associada uma *tabela de probabilidades condicionais*, $P(A | pa(A))$. Os nós pais de A, $pa(A)$, podem ser outros nós de incerteza ou nós de decisão. Se ele não possuir nó pai, então a sua tabela de probabilidade é substituída pelas probabilidades a priori $P(A)$ da variável em questão (neste caso, é considerada uma variável de evidência) e possuem um conjunto finito de estados mutuamente exclusivos.
- **Nós de decisão (D)** são pontos de escolha de ações (retângulos). Seus nós pais podem ser outros nós de decisão ou nós de incerteza e tem um conjunto finito de estados mutuamente exclusivos. Existe um caminho direcionado ligando todos os nós de decisão.
- **Nós de utilidade (U)** são funções de utilidade (losangos), portanto, para cada nó (U) está associada uma *função real* sobre $pa(U)$ e possui uma tabela contendo descrição da utilidade do decisor como função das variáveis associadas aos seus nós pais. Seus pais podem ser nós de decisão ou nós de incerteza, eles não têm filhos e nem estados.

Os elementos de um diagrama de influência são representados por



Um diagrama de influência fornece representações compactas dos problemas de decisão enquanto suprime muito dos detalhes, sendo, portanto, ideal para se obter uma visão geral, especialmente de problemas complexos.

Exemplo 2: Uma fábrica de componentes de automóvel - Pethold - esta enfrentando uma nova demanda por um de seus produtos. Um dos diretores e quatro executivos se reúnem para considerar formas alternativas de lidar com este eventual aumento da demanda.

Após algumas discussões concluem por duas ações alternativas capazes de atender á nova demanda:

- a_1 - comprar novos equipamentos (NvEqui)
- a_2 - contratar horas extras (HrExt)
- a_3 - manter nível de produção atual (NvAtua)

O Diretor não admite subcontratar outro fornecedor por questões estratégicas. Além disto, para simplificar, não há expectativas de variações nos preços. Após discutirem o que aconteceria sob cada uma das alternativas e decidirem trabalhar com um horizonte de planejamento de um ano, o pessoal de marketing julgou que a demanda, a se manter a tendência atual, poderia subir uns 15% (Alta), mas não exclui a possibilidade de uma queda de 5% (Baixa) caso o mercado se torne sofrível. As chances relativas com relação a esses dois estados são de 3:2 a favor do crescimento das vendas, ou seja, a probabilidade de Alta é igual a 0.6.

Neste momento solicitaram ao pessoal de contabilidade que levantasse os custos de cada uma das alternativas. Após várias idas e vindas, envolvendo os gerentes de produção, de pessoal e financeiro, responsáveis pelos custos de material e equipamentos, de salários e pelos custos financeiros, respectivamente, chegaram aos números da tabela 4.2.

Tabela 4.2: Custos em unidades monetárias (u.m)

Decisões	Estados	
	Alta	Baixa
NvEqui	220	130
HrExt	210	150
NvAtua	170	150

Essa tabela contém vários ingredientes, por exemplo, NvEqui, HrExt e NvAtua pertencem ao espaço das ações (decisões), enquanto Alta e Baixa compõem os estados da natureza. A tabela 4.2 representa a função objetivo (custos, perdas, ganhos, utilidades, etc).

Dado o volume de informações, pode-se organizar um diagrama de influência, como mostra abaixo a figura 4.3a.

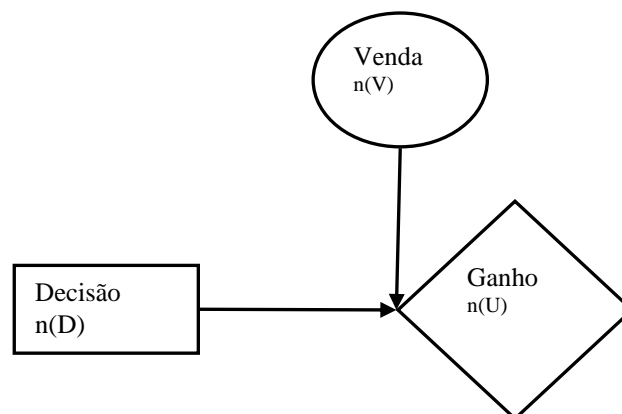


Figura 4.3a: Diagrama de influência do exemplo 1.

Utilizando um pacote de análise de decisões para introduzir os dados, obtém-se uma árvore de decisão (figura 4.3b) e resolver o problema. Em um problema de decisão, existe uma relação biunívoca entre a árvore de decisão e o diagrama de influência, sendo o segundo de mais fácil elaboração em problemas relativamente complexos.

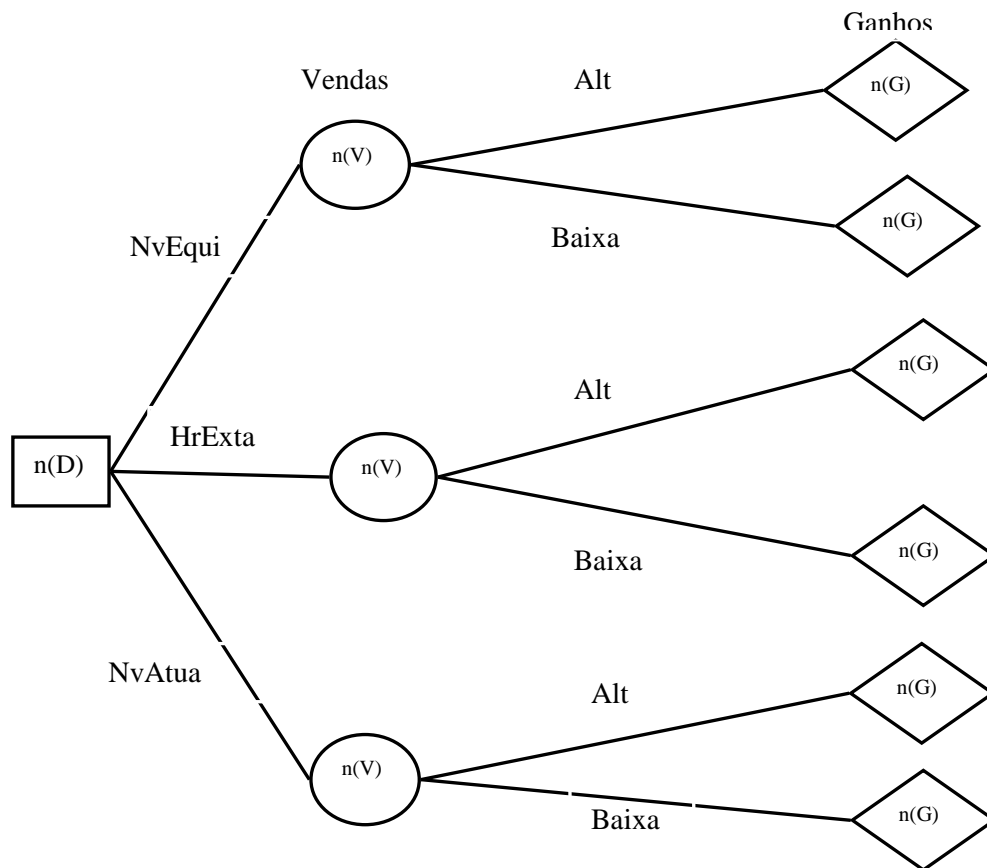


Figura 4.3b: Árvore de Decisão do exemplo 1.

Exemplo 3: O governo oferece a uma companhia de petróleo a opção de perfurar um campo “A” ou perfurar um campo “B”. A probabilidade de existir petróleo no campo “A” é de 0.4, e no campo “B” é de 0.2 (esses dois eventos são independentes). O ganho líquido esperado se o petróleo for extraído é de \$77 milhões para o campo “A”, e \$195 milhões para o campo “B”. As possíveis decisões da companhia são:

- i. Não perfurar nenhum dos campos;
- ii. Perfurar o campo “A” ou o campo “B”;
- iii. Pagar para realizar uma investigação prévia em um dos campos e segundo o resultado dessa investigação escolher entre as opções i. e ii.

Deve ter observado que o resultado da investigação prévia não é totalmente preciso. Assim, se existir petróleo em um dos campos o resultado da investigação aconselhará perfurar com probabilidade de 0.8, se não o resultado da investigação aconselhará perfurar com probabilidade 0.4. O custo da investigação é de \$6 milhões, e o custo de perfurar qualquer dos campos é de \$31 milhões.

A figura 4.3c, ilustra o digrama de influência do exemplo anterior (três).

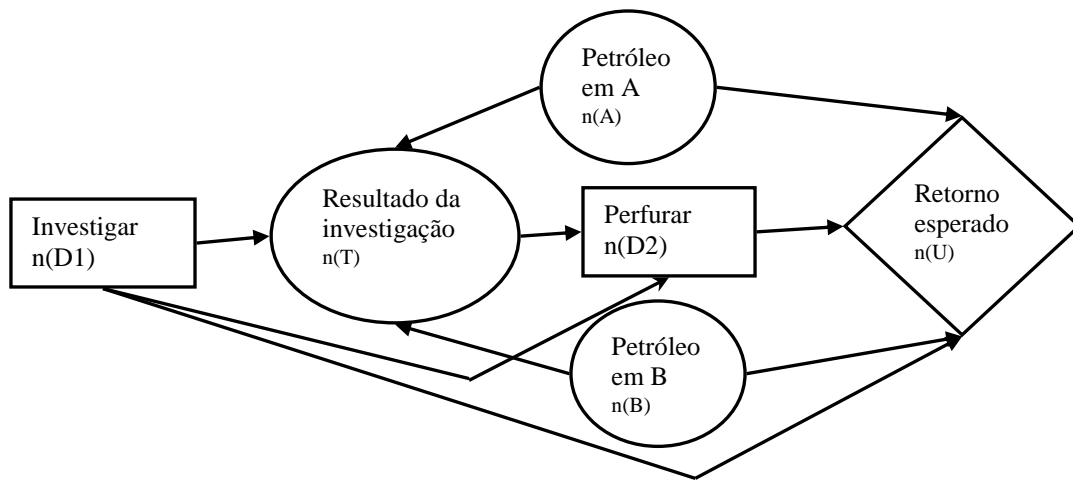


Figura 4.3c: Diagrama de influência do exemplo 2.

O diagrama de influência foi desenvolvido para substituir árvores de decisão convencional em modelagem e resolução de problemas de decisão do mundo real.

Comparativamente, o diagrama de influência oferece vantagens na formulação compacta e intuitiva, fácil avaliação numérica, representação efetiva das independências entre as variáveis sobre as árvores.

Esses fatores nos últimos anos contribuíram para a ampla utilização dos diagramas de influência como uma ferramenta para representação e análise de problemas de decisão complexos.

Capítulo 5

Grafos de Modelos Estatísticos

A modelagem por grafos é uma forma de análise multivariada que utiliza grafos para representar modelos.

De certo modo, modelagem com DAGs é simples. Uma vez que a densidade condicional $f(V_j | V_{j-1} \dots V_1)$ pode ser livremente especificada, e qualquer modelo univariado adequado pode ser usado. Isto é, para cada j , podemos modelar a dependência " V_i " sobre uma variável a priori $V_1 \dots V_{j-1}$ usando algum modelo, em que $V_1 \dots V_{j-1}$ são incluídos como covariáveis; entretanto se apenas um subconjunto é incluído, então " V_i " depende apenas deste subconjunto.

Há evidentemente, uma enorme variedade de modelos univariados ou multivariados que podem ser aplicados, entre eles podemos citar modelos lineares generalizados (McCullagh and Nelder, 1989) e modelos de regressão para variáveis ordinais (Agresti, 1984) citados por Edwards (1995). Tipos diferentes de modelos podem ser usados para cada medida.

Um aspecto que se merece enfatizar, é que a escolha do modelo em cada etapa é bastante independente da escolha do modelo a todos os outros passos.

Por exemplo, considere um vetor aleatório k -dimensional $X = (X_1, X_2, \dots, X_k)$ e um grafo $G = (V, A)$. Um modelo para X é uma família de distribuições de probabilidade para X , que satisfaz as restrições de independência condicionais para pares de variáveis em G .

Quando a distribuição é normal multivariada então falamos do modelo Gaussiano. Neste caso, temos que a independência condicional é equivalente à especificação de zeros na matriz de covariâncias inversas.

5.1. Análise Multivariada

5.1.1. Variáveis Discretas – Modelo Log linear

Modelo log-linear é uma técnica de análise de dados multivariada baseada na distribuição Multinomial utilizada para detectar o padrão de dependência entre variáveis categóricas, que é bastante útil em trabalhos na área de Saúde Pública e Epidemiologia, na qual é comum se ter tabelas de contingência complexas, com grande número de variáveis.

Edwards (1995) e Whittaker (1990) mostram que existem subclasses de modelos log-lineares hierárquicos, especificados pela parametrização da função densidade em termos dos coeficientes da expansão log-linear, os termos " u ".

Os modelos log-lineares são assim chamados porque eles usam modelos lineares para logaritmos da contagem de elementos esperados. Para motivar este assunto, consideremos N observações discretas A, B e C. Tomamos 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 $1, \dots, \#C$.

E similarmente, escrevemos a probabilidade da célula (j, k, l) da tabela como p_{jkl} e o valor esperado $m_{jkl} = N \times p_{j,k,l}$, com N sendo o número de observações.

Um modelo simplificado para uma tabela de três entradas escreve o logaritmo da probabilidade de cada célula como sendo

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

Em que os u 's são os parâmetros, também chamados de termos de interação.

De referir que o modelo da equação acima, A, B e C são completamente independentes. Para identificar um modelo, uma simples formula pode ser usada.

Esta consiste de uma lista de termos chamados geradores, que corresponde à interação maximal no modelo.

Para o modelo independente, a fórmula é A, B e C e o grafo independente correspondente é mostrado na figura abaixo (figura 5.1.1a).

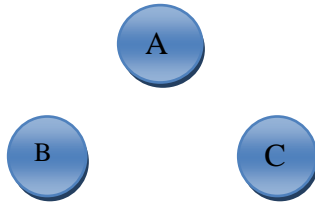


Figura 5.1.1a: Modelo grafo para variáveis A, B e C independentes.

Um modelo mais complexo pode ser escrito como

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

Desde que $u_{jk}^{AB} + u_{jl}^{AC}$ sejam termos de interação maximal, o modelo será AB, AC. O modelo pode ser reformulado em termo de probabilidade de elementos como:

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

Portanto, a formula acima revela que B e C são condicionalmente independentes dado A, e denotamos por $B \perp\!\!\!\perp C | A$. O grafo para este modelo é representado na figura 5.1.1b.

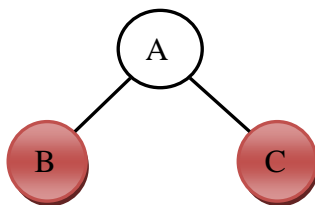


Figura 5.1.1b: Modelo grafo para variáveis A, B e C onde $B \perp\!\!\!\perp C | A$

O modelo log-linear geral para três fatores (modelo fórmula ABC) pode ser escrito como:

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

O modelo anterior “AB, AC” pode ser considerado como um caso especial do modelo geral definido pela posição $u_{kl}^{BC} = u_{jkl}^{ABC} = 0$, 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 (figura 5.1.1a) é caso especial do modelo geral com $u_{jk}^{AB} = u_{jl}^{AC} = u_{kl}^{BC} = u_{jkl}^{ABC} = 0$.

Muitas das vezes, a atenção é restrita para modelos log-lineares *hierárquicos*. Aqui o termo hierárquico indica que se um termo é zero, então todos de ordem superior relativos são também nulos. Por exemplo, se no nosso conjunto $u_{jk}^{AB} = 0$, mas admitimos não zeros no termo de interação de três fatores u_{jkl}^{ABC} , então este poderia definir um modelo não-hierárquico. Os modelos não hierárquicos são de difícil interpretação, razão pela qual, na prática, eles não são muito usados.

Para o modelo AB, AC, que é definido pela posição $u_{kl}^{BC} = 0$, e daí $u_{jkl}^{ABC} = 0$, 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 não envolve B. Então, do critério de fatoração, segue que $B \perp C | A$. 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 igual a 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 entre 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 independente desde que u_{kl}^{BC} seja zero. O exemplo simples de um modelo grafo não hierárquico é AB, BC, AC, o qual o conjunto de interação u_{jkl}^{ABC} de três fatores é igual a zero.

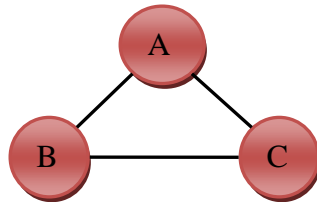


Figura 5.1.1c: Modelo grafo não hierárquico AB, BC e AC

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

Para ajudar a entender o exemplo do grafo e modelos hierárquicos, vamos considerar uma tabela de três entradas de Fienderg (1980) citado por Edwards (1995), sobre comportamento de ninhos de lagartos. A tabela 5.1 consiste de dados de comportamentos nos ninhos de duas espécies de lagartos.

Para amostra de 164 lagartos, são dadas as seguintes variáveis: 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.

Tabela 5.1: Dados de comportamento dos ninhos de duas espécies de lagartos.

Especie	Diametro Ninho	Altura Ninho	Altura Ninho (< 4.75)
		(> 4.75)	
Anoli	≤ 4	32	86
Anoli	> 4	11	35
Distichus 2	≤ 4	61	73
Distichus 2	> 4	41	70

Com o resultado de teste estatístico efetuado por Edwards (1995, pp19), é necessário excluir a aresta BC do grafo completo, resultando no modelo AB, AC, e com ajuda da *deviance*, vemos que este é um modelo aceitável. O grafo resultante deste modelo é dado na figura 5.1.1d.

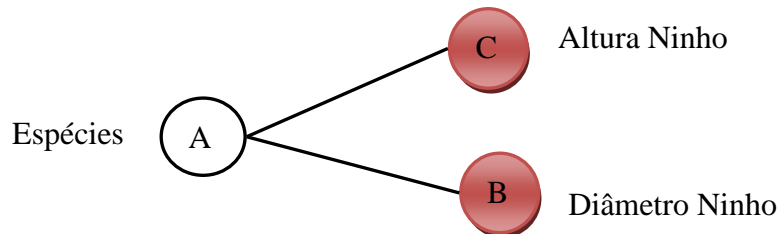


Figura 5.1.1d: Grafo resultante no modelo AB, AC

O grafo da figura 5.1.1d revela que o diâmetro e a altura dos ninhos são condicionalmente independentes dada a espécie do lagarto, isto é, $B \perp C | A$.

5.1.2. Variáveis Contínuas – Modelo Gaussiano

Em geral, as técnicas de análise multivariada são baseadas nas informações da matriz de covariância da variância. Esta matriz é quadrada e simétrica com várias linhas e colunas de acordo com o número de variáveis em análise. Na diagonal são as variâncias de cada variável e o resto das células tem as covariâncias.

Para os casos de modelo grafo ser Gaussiano ou em distribuição normal multivariada as independências condicionais entre variáveis são equivalentes à especificação de zeros na matriz de covariâncias inversa.

Portanto, sob a suposição de normalidade multivariada, valor zero numa célula na matriz inversa de covariância significa que as duas variáveis correspondentes são condicionalmente independentes dado as demais variáveis.

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

- (i) Estimar a matriz de covariância $V = \text{var}(X)$ pela matriz de covariância amostral, S .
- (ii) Calcular a inversa S^{-1} , a inversa da matriz de correlação amostral. Os elementos da diagonal são interpretados em termos de variâncias parciais.
- (iii) Escalonar S^{-1} até obter uma diagonal unitária e calcular as correlações parciais amostrais, $\text{cor}(X_i, X_j | \text{resto})$.
- (iv) 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.

Portanto, temos que os vetores aleatórios normais X_a e X_b são independentes se, e somente se,

- (i) $\text{cov}(X_a, X_b) = 0$, parametricamente expressado, $V_{ab} = 0$; ou
- (ii) $D_{ab} = 0$, onde D é a inversa da variância.

Por outro lado, os vetores aleatórios normais X_b e X_c são condicionalmente independentes dado X_a se, e somente se,

- (i) $\text{cov}(X_b, X_c | X_a) = 0$; ou
- (ii) O bloco de inversa da variância $D_{bc} = 0$.

Em particular, se X_b e X_c são ambos unidimensionais, então $X_i \perp\!\!\!\perp X_j | X_{k \setminus \{i,j\}}$ se, e somente se, $d_{ij} = 0$

Por exemplo, considere o vetor aleatório X de dimensão quatro, $k = 4$, e um modelo grafo Gaussiano para X é tal que a inversa da variância $D = \text{var}(X_1, X_2, X_3, X_4)^{-1}$ 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 será dado pela figura 5.1.2a

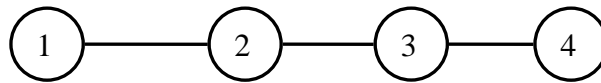


Figura 5.1.2a: grafo não orientado com quatro vértices

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

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

Exemplo 4: trata-se de um exemplo numérico (variáveis contínuas) cujos dados, segundo Whittaker (1990), refere-se a uma amostra de notas de estudante em cinco disciplinas: mecânica, vetores, álgebra, análise e estatística. Todas as variáveis são medidas na mesma escala, e não há valores desconhecidos ou omitidos.

Os referidos dados podem ser encontrados em Whittaker (1990) pagina dois, e forneceram a matriz de variância abaixo apresentada.

Tabela 5.2: Matriz de variância de 5 disciplinas

	mec	vet	Alg	anal	est
mec	302.29				
vet	125.78	170.88			
alg	100.43	84.19	111.60		
anal	105.07	93.60	110.84	217.88	
est	116.07	97.89	120.49	153.77	294.37

Uma matriz de correlação é obtida da matriz de variâncias pelo escalonamento de linhas e colunas de tal forma que as entradas diagonais sejam todas unitárias. Assim, segue abaixo a matriz de correlação resultante deste escalonamento.

Tabela 5.3: Matriz de correlações de 5 disciplinas

	mec	vet	Alg	anal	est
mec	1.0				
vet	0.55	1.0			
alg	0.55	0.61	1.0		
anal	0.41	0.49	0.71	1.0	
est	0.39	0.44	0.66	0.61	1.0

Na matriz de correlações, todos os valores são positivos, o que pode levar a dizer que um bom estudante tende a obter boas notas em todas as matérias, enquanto um mau estudante tende em tirar notas baixas em todas as matérias. O grau mais alto de correlação (0.71) foi obtido entre álgebra e análise, enquanto o mais baixo (0.39) entre a mecânica e estatística.

Na prática não é bem conhecida a interpretação da matriz inversa da variância e de correlações. Todavia, o que motiva o cálculo da matriz inversa da variância é se um

elemento da matriz inversa da variância for igual a zero indica que as variáveis correspondentes são condicionalmente independentes dado as variáveis restantes.

Entretanto, a prática mostra que é mais fácil interpretar uma correlação do que uma covariância.

Assim, a matriz inversa de correlações é calculada

Tabela 5.4: Matriz inversa de correlações das 5 disciplinas

	mec	vet	Alg	anal	est
mec	1.60				
vet	-0.56	1.80			
alg	-0.51	-0.66	3.04		
anal	0.00	-0.15	-1.11	2.18	
est	-0,04	-0.04	-0.86	-0.52	1.92

Cada elemento diagonal da matriz de correlações inversa está relacionado com a proporção de variação da variável correspondente explicada 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.

Entretanto, escalonando a inversa da matriz de correlações para obter a unidade na diagonal, temos

Tabela 5.5: Matriz inversa de correlações escalonada das 5 disciplinas

	mec	vet	Alg	anal	est
mec	1				
vet	-0.33	1			
alg	-0.23	-0.28	1		
anal	0.00	-0.08	-0.43	1	
est	-0.02	-0.02	-0.36	-0.25	1

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

Ao contrário da matriz de correlação original das notas, há uma estrutura identificável nesta matriz de correlação inversa escalonada. Os elementos do lado esquerdo abaixo da diagonal são todos próximos a zero, o que sugere que a matriz pode ser aproximada pela estrutura abaixo, onde os valores diferentes de zeros são representados por “*”.

Tabela 5.6: Estrutura da matriz inversa de correlações escalonada

	mec	vet	Alg	anal	est
mec	*				
vet	*	*			
alg	*	*	*		
anal	0	0	*	*	
est	0	0	*	*	*

Desta forma, a partir da matriz de correlação inversa escalonada podemos construir um grafo, onde os vértices representaram as matérias e vai existir uma aresta entre dois vértices se o valor na sua interseção é diferente de zero (figura 5.1.2b).

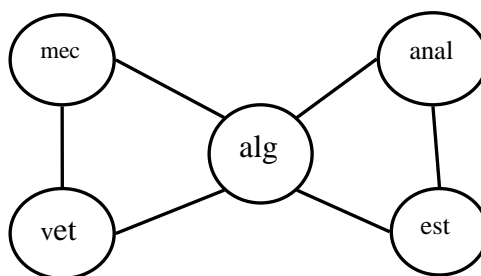


Figura 5.1.2b: grafo de independência condicional de cinco disciplinas

A figura 5.1.2b é grafo de independência condicional para distribuição conjunta normal multivariada de variáveis aleatórias referente às notas das cinco disciplinas.

Por exemplo, na matriz de correlação inversa escalonada, a correlação parcial entre mecânica e análise é zero o que indica que estas variáveis são condicionalmente independentes dadas às restantes variáveis (vetores, álgebra e estatística).

Mas como no grafo, álgebra separa mecânica e análise, conclui-se que estas duas variáveis são condicionalmente independente dado álgebra somente.

Por outro lado, verifica-se que as propriedades de Markov estão inclusas no grafo (5.1.2c):

- Propriedade de Markov para pares de variáveis (parewise)

$\text{mec} \perp\!\!\!\perp \text{anal} \mid (\text{alg}, \text{vet}, \text{est});$

$\text{vet} \perp\!\!\!\perp \text{est} \mid (\text{alg}, \text{mec}, \text{anal});$

$\text{mec} \perp\!\!\!\perp \text{est} \mid (\text{alg}, \text{anal}, \text{vet})$

$\text{vet} \perp\!\!\!\perp \text{anal} \mid (\text{alg}, \text{mec}, \text{est}).$

- Propriedade de Markov Local

$\text{mec} \perp\!\!\!\perp (\text{anal}, \text{est}) \mid (\text{alg}, \text{vet});$

$\text{vet} \perp\!\!\!\perp (\text{anal}, \text{est}) \mid (\text{alg}, \text{mec});$

$\text{anal} \perp\!\!\!\perp (\text{mec}, \text{vet}) \mid (\text{alg}, \text{est});$

$\text{est} \perp\!\!\!\perp (\text{mec}, \text{vet}) \mid (\text{alg}, \text{anal})$

- Propriedade de Markov global

Consideremos os subconjuntos $b = \{mec, vet\}$, $c = \{anal, est\}$ e $a = \{alg\}$, desta forma, temos que $X_b \perp\!\!\!\perp X_c \mid X_a$.

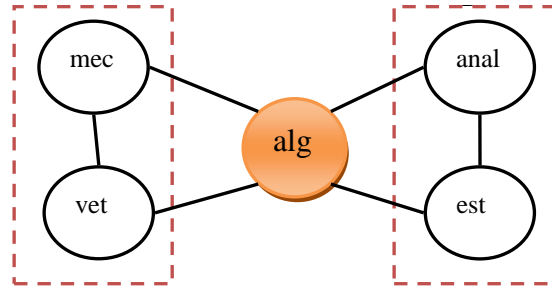


Figura 5.1.2c: grafo ilustrando a propriedade global de Markov nas cinco disciplinas

5.2. Análise de Regressão

Regressão é uma técnica que permite explorar e inferir a relação de variável dependente (variável de resposta) com variáveis independentes específicas (variáveis explicatórias).

A análise de regressão pode ser usada como um método descritivo da análise de dados (como por exemplo, o ajustamento de curvas) sem serem necessárias quaisquer suposições acerca dos processos que permitiram gerar os dados.

Regressão designa também uma equação matemática que descreva a relação entre duas ou mais variáveis.

É importante lembrar que todo modelo é uma simplificação da realidade. Desta forma, todos os modelos estatísticos ou probabilísticos apresentam um componente de erro. Isto indica que, mesmo o modelo tendo um bom poder de explicação, ele sempre incorrerá e esse erro deve ser minimizado.

A regressão linear é chamada “linear” porque se considera que a relação da resposta às variáveis é uma função linear de alguns parâmetros. Os modelos de regressão que não são uma função linear dos parâmetros são chamados de modelos de regressão não linear.

De referir, que um parâmetro é uma medida descritiva (média, variância, proporção, etc.) dos valores x_1, x_2, x_3, \dots , associados à uma população. Enquanto que Estatísticas são medidas associadas à amostras.

Em análise de regressão existem dois tipos de variáveis: variáveis independentes (explicativas) e variáveis dependentes ou respostas (explicada).

5.2.1. Modelo de regressão Linear simples

Modelo de regressão Linear simples ou simplesmente regressão linear simples é um método para se estimar o valor esperado de uma variável resposta “Y”, dados os valores de alguma outra variável explicativa ou independente “X”.

O grafo abaixo que ilustra uma relação em um modelo de regressão linear simples, onde o vértice “e” representa a componente de erro.

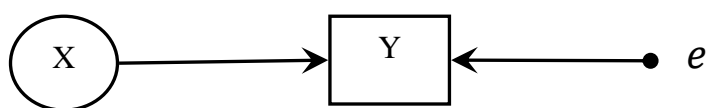


Figura 5.2.1: grafo para regressão linear simples

O grau de correlação entre duas variáveis continua é quantificado por uma medida chamada de coeficiente de correlação de Pearson, onde do qual, também é ilustrada a direção da correlação em causa (se positiva ou negativa).

O coeficiente de correlação de Pearson é obtido dividindo-se a covariância de X e Y pelo produto dos respectivos desvios-padrão de ambas as variáveis, isto é:

$$\rho = \text{cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

Este coeficiente, normalmente representado por r e seu valor varia entre -1 e 1.

- $r = 1$ significa que as duas variáveis têm uma correlação linear perfeita positiva.
- $r = -1$ significa que as duas variáveis têm uma correlação linear perfeita negativa, isto é, se uma aumenta, a outra tende sempre a diminuir.
- $r = 0$ significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma dependência não linear. Assim, o resultado $r = 0$ deve ser investigado por outros meios.

Para estimar o valor esperado em uma regressão linear simples, usa-se uma equação que determina a relação entre ambas as variáveis.

$$E(Y) = \beta_0 + \beta_1 x_i + e$$

Um valor de $\beta_1 = 0$ isso implica que a variável Y é independente a variável X, isto é, $Y \perp X$ ou $X \perp Y$. Portanto, neste caso não existe uma aresta ou arco que possa unir o vértice Y e X.

Em muitas situações a variável dependente em que estamos interessados y pode ser afetada por mais de uma variável explicativa (x_i). Em tais casos, devemos aplicar um modelo chamado de **regressão múltipla**.

5.2.2. Regressão Múltipla

A análise de *regressão múltipla* envolve o uso de duas ou mais variáveis explicativas. Conceitualmente, a regressão múltipla nada mais é que uma extensão da regressão linear simples.

Esse modelo pode ser representado em forma do seguinte grafo:

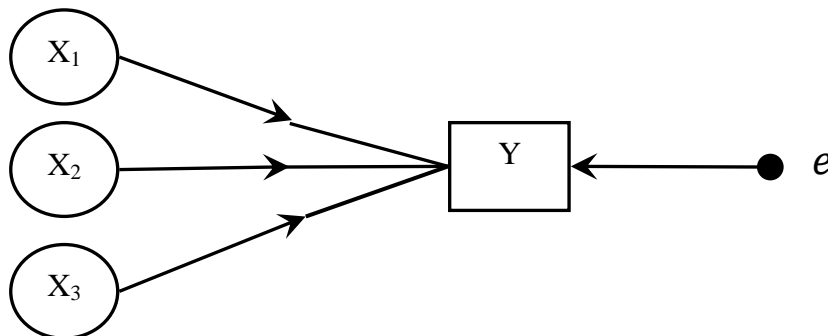


Figura 5.2.2a: grafo para regressão linear múltipla.

Para estimar o valor esperado em uma regressão linear múltipla, usa-se uma equação que determina a relação entre ambas as variáveis.

$$E(Y) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e$$

O conceito de independência condicional também aplicado na análise de regressão múltipla, de modo que, uma hipótese de que o i -ésimo coeficiente de regressão $\beta_i = 0$ é equivalente a hipótese de que a variável Y (explicada) e a variável explicativa X_i são condicionalmente independentes dado as restantes variáveis no modelo.

Portanto, a equação de regressão deste modelo não incluirá a variável com $\beta_i = 0$.

O grafo abaixo ilustra uma relação entre uma variável Y e as variáveis explicativas X_i (X_1 , X_2 e X_3).

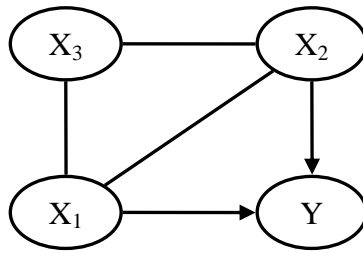


Figura 5.2.2b: grafo para regressão linear múltipla com três variáveis explicativas.

A equação de regressão correspondente ao grafo anterior será $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, portanto, o coeficiente β_3 é igual a zero ($\beta_3 = 0$).

Assim, $X_3 \perp\!\!\!\perp Y \mid (X_1, X_2)$.

A interpretação dos coeficientes de regressão na regressão múltipla é muitas vezes fonte comum de confusão. A equação de regressão linear simples representa uma linha, enquanto a equação de regressão múltipla representa um plano (em caso de duas variáveis preditoras) ou um hiperplano (em casos de mais de dois indicadores).

5.2.3. Regressão Multivariada

As técnicas de análise de regressão linear têm sido amplamente utilizadas tanto na literatura como em problemas práticos. Nos modelos de regressão apresentados anteriormente, nos restringimos predominantemente a modelos em que a variável resposta (dependente) é única associada a um ou conjunto de variáveis preditoras (explicativas).

Na situação de modelo de **regressão Multivariada**, a variável resposta (dependente) é multivariada, ou seja, é de forma $Y = (Y_1, \dots, Y_p)$, sendo que as variáveis Y_1, \dots, Y_p são medidas na mesma unidade amostral e, como no caso univariado, associadas a p conjuntos de variáveis preditoras (explicativas).

Para a definição do modelo, vamos supor o caso em que dispomos de p variáveis respostas (Y_1, \dots, Y_p).

A cada uma dessas p variáveis está associado um grupo de variáveis explicativas X_{11}, \dots, X_{1k_1} .

Procedendo dessa forma para as demais variáveis respostas, X_{p1}, \dots, X_{pk_1} é um conjunto de variáveis explicativas associadas à variável resposta Y_p .

Um grafo para um modelo de regressão Multivariada com três variáveis explicativas e duas explicadas é apresentado na figura 5.2.3.

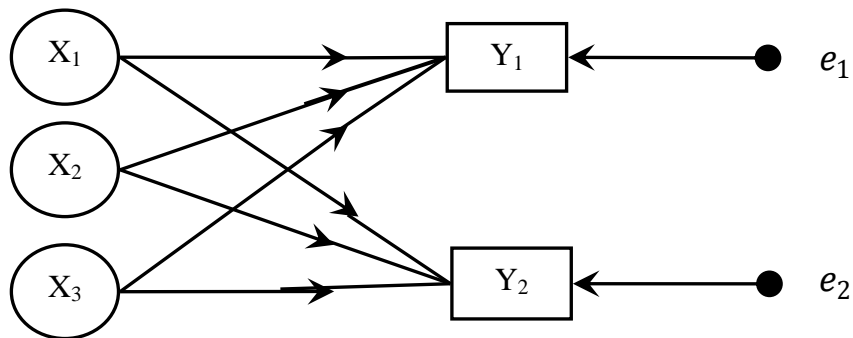


Figura 5.2.3: grafo para regressão linear multivariada.

5.3. Equações estruturais (ou equações simultâneas)

Ainda dentro da análise multivariada, há outro conceito que frequentemente em determinadas condições é usada para análise de dados multivariados que são as *equações estruturais lineares*.

Equações estruturais pode ser vista como sendo particularmente útil quando uma variável dependente em subseqüentes relações de dependência. Isto é, a preocupação nesta técnica é com a ordem das variáveis. Na regressão multivariada tradicional temos

que X influencia Y ; em modelação de equações estruturais, X influencia Y e Y influencia Z ou por outra X influencia Y_i e Y_i influencia Y_j .

Existem dois tipos de variáveis em um modelo de equações estruturais (Farias e Santos (2000), citando Loehlin, 1998): variáveis endógenas e variáveis exógenas. Os valores das variáveis endógenas são explicados por uma ou mais variáveis exógenas do modelo. Os valores das variáveis exógenas são assumidos como dados, isto é, o modelo não tenta explicá-los.

Esta distinção é similar à feita entre variáveis dependentes (endógenas) e independentes (exógenas) da análise de regressão. Entretanto, em um modelo de equações estruturais, uma variável pode ser tanto dependente quanto independente.

A regra é esta: se uma variável é dependente em alguma parte do modelo, então ela é endógena (Farias e Santos (2000), citando Klem, 1995).

Por exemplo, um modelo de equações estruturais com três variáveis explicativas (X) e duas explicadas (Y) pode ser representado em forma de grafos de acordo com a figura abaixo (figura 5.3a).

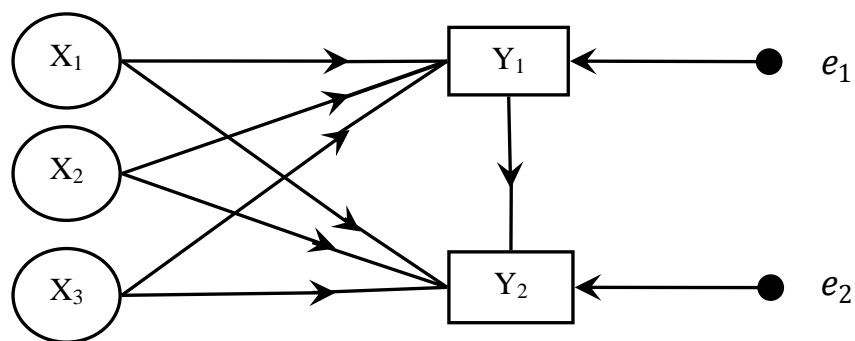


Figura 5.3a: grafo para equações estruturais ou simultâneas.

Em outras vezes, de acordo com Farias e Santos (2000), a modelagem de equações estruturais tem sido rotulada de modelagem de caminhos (*path modeling*), análise de caminhos (*path analysis*), e ou análise de variáveis latentes de equações estruturais (*latent variable analysis of structural equations*).

Hair *et al.* (1998) afirma que *Variável latente* não é nada mais que uma variável que não pode ser diretamente medida, mas pode ser representada ou medida por outras variáveis (indicadores). Por exemplo, a atitude de uma pessoa em relação a um produto jamais pode ser medida com precisão a ponto de não haver incerteza, mas fazendo-se varias perguntas podemos avaliar muitos aspectos da atitude dessa pessoa.

Em contribuição, as respostas a tais questões fornecem uma medida razoavelmente precisa da variável latente (atitude) para um individuo.

Para Farias e Santos (2000), citando Klem (1995) a análise de caminhos foi usada pela primeira vez em 1918 por Sewall Wright, um geneticista que resolveu equações estruturais ou simultâneas para desvendar influências genéticas através de gerações.

Ele desenvolveu um modelo unidirecional, em que as causas e as saídas eram conhecidas, e a causalidade era estipulada em uma única direção, sem *feedback*.

Este modelo é conhecido como unidirecional ou recursivo. É o único que pode ser apropriadamente chamado de análise de caminhos (Farias e Santos (2000) citando Maruyama, 1998).

Um exemplo de um diagrama de caminhos (grafo) referente a uma análise de equações estruturais com algumas variáveis latentes é mostrado na figura 6.3b. Os dados são de um estudo sobre o desempenho e satisfação; especificamente, estudo de Bagozzi (1980) citado por Dillon (1984) que formulou um modelo de equações estruturais para o estudo de relação (conexão) entre desempenho e satisfação em uma potencia de vendas industriais.

O modelo foi desenhado para responder perguntas como: "existe uma relação entre desempenho e satisfação? Será que o desempenho influencia a satisfação, ou a satisfação influencia o desempenho?" (Bagozzi 1980, PP. 65, citado por Dollin (1984).

O modelo adotado por Bagozzi consiste em três variáveis exógenas latentes (ξ_1, ξ_2 e ξ_3) e duas variáveis endógenas latentes (η_1 e η_2), onde

ξ_1 = *motivação realização*

ξ_2 = *auto – estima em tarefas específicas*

ξ_3 = *inteligencia verbal*

η_1 = *desempenho*

η_2 = *satisfação no trabalho*

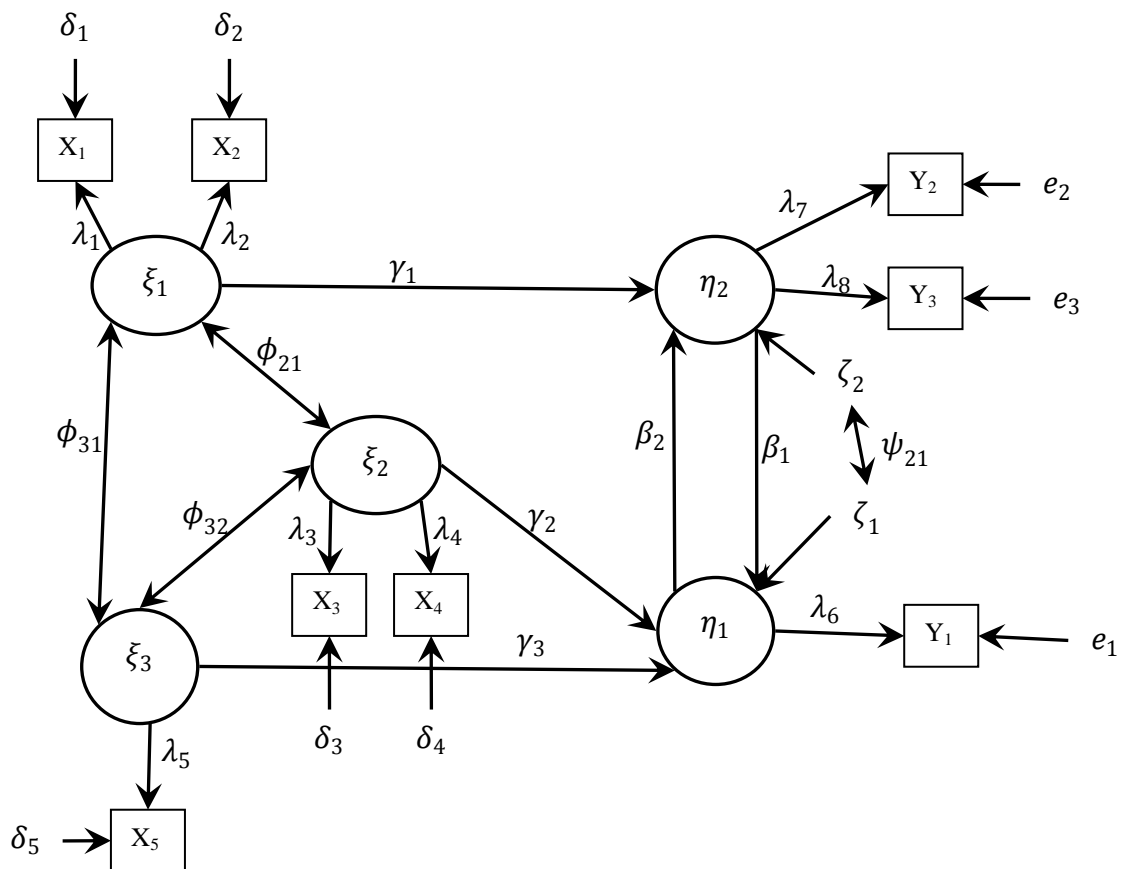


Figura 5.3b: diagrama de caminhos para o desempenho e a satisfação.

As variáveis ζ_1 e ζ_2 são erros da equação (residual), e as variáveis denotadas por X são medições observadas para as variáveis latentes ξ . Da mesma forma, as variáveis denotadas por Y são as medições observadas para a variável latente η , β é o efeito de uma variável dependente sobre outra variável dependente, o coeficiente γ é o efeito de uma variável independente sobre outra variável dependente, ϕ é a correlação entre as variáveis independentes não observadas, λ é o coeficiente de regressão entre as

variáveis observadas e as variáveis não observadas. Ambos δ e ε denotam erros que correspondem à confiabilidade de medição.

Com base nos dados coletados referentes às variáveis envolvidas no modelo da figura acima, obteve-se uma matriz de coeficientes de correlações de Pearson (tabela 5.7)

Tabela 5.7: matriz de correlação para valores observados, fonte, Dollin (1984).

Variable	Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4	X_5
Performance, Y_1	1.000							
Job satisfaction 1, Y_2	.418	1.000						
Job satisfaction 2, Y_3	.394	.627	1.000					
Achievement motivation 1, X_1	.129	.202	.266	1.000				
Achievement motivation 2, X_2	.189	.284	.208	.365	1.000			
Self-esteem 1, X_3	.544	.281	.324	.201	.161	1.000		
Self-esteem 2, X_4	.507	.225	.314	.172	.174	.546	1.000	
Verbal intelligence, X_5	-.357	-.156	-.038	-.199	-.277	-.294	-.174	1.000

Source: Bagozzi (1980).

De acordo com o modelo, ξ_1 é medido por dois indicadores (X_1 e X_2), ξ_2 é medido por dois indicadores (X_3 e X_4) e a inteligência verbal (ξ_3) é medido por apenas um indicador (X_5). Assume-se ainda que o desempenho (η_1) é medido por um único indicador (Y_1) e a satisfação no trabalho (η_2) é medido por dois indicadores (Y_2 e Y_3).

O modelo retratado na figura 5.3b sugere que a variável motivação na realização (ξ_1) afeta diretamente na satisfação de trabalho (η_2) e indiretamente ao desempenho (η_1) via a satisfação no trabalho (η_2). A auto-estima em tarefas específicas (ξ_2) e a inteligência verbal (ξ_3) afetam diretamente o desempenho (η_1) e indiretamente a satisfação no trabalho (η_2) via desempenho (η_1).

Ainda neste modelo, o desempenho afeta diretamente a satisfação no trabalho e por outro lado, a satisfação afeta diretamente o desempenho. Entretanto, modelos com efeitos recíproco nas variáveis dependentes são chamados de *modelos não recursivos*, caso contrário são chamados *recursivos* (Dollion, 1984). De referir que para o primeiro caso (*modelos não recursivos*) não é necessariamente que $\beta_1 = \beta_2$.

Se nós supusermos que o desempenho (η_1) afeta diretamente a satisfação no trabalho (η_2), mas não vice versa, então o modelo da figura 5.3b seria um modelo recursivo; isto é, β_1 seria definido como zero. Portanto, este modelo (figura 5.3b) corresponde à seguinte sistema de equações estruturais:

$$\eta_1 = \beta_1\eta_2 + \gamma_2\xi_2 + \gamma_3\xi_3 + \zeta_1$$

$$\eta_2 = \beta_2\eta_1 + \gamma_1\xi_1 + \zeta_2$$

Onde ζ_1 e ζ_2 são resíduos, isto é, erros na equação. Note que na mesma figura o ψ_{21} fornece a correlação entre ζ_1 e ζ_2 .

Considere agora que temos como objetos de estudo a relação de cinco variáveis, concretamente o mesmo caso da figura 5.3b, com exceção que neste caso apenas estamos lidando com variáveis observáveis, em outras palavras, a implicação é que os indicadores individuais X_1, X_2, X_3, Y_1 e Y_2 são perfeitos indicadores das suas respectivas variáveis latentes.

Temos *indicador simples da motivação na realização* (X_1),
indicador simples da auto – estima em tarefas específicas (X_2),
indicador simples da inteligencia verbal (X_3),
indicador simples do desempenho (Y_1) e *um indicador simples da satisfação no trabalho* (Y_2), o que resulta no grafo da figura 5.3c.

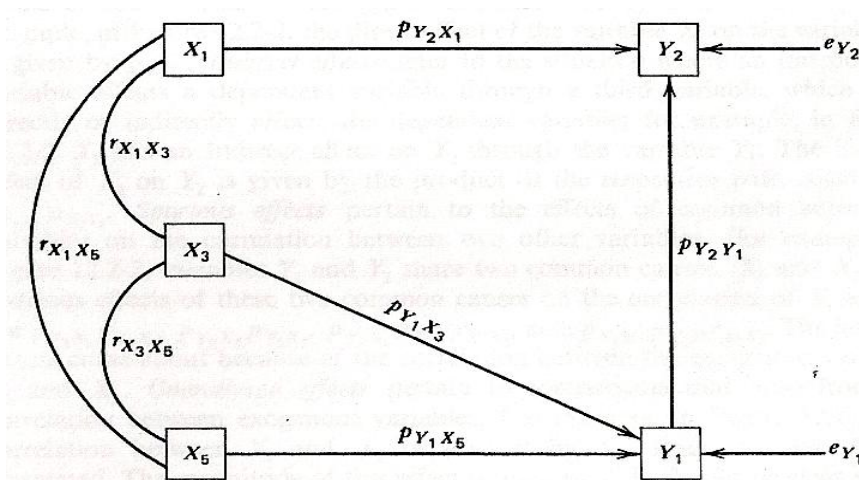


Figura 5.3c: Grafo mostrando a relação entre três variáveis independentes e duas dependentes.

Portanto, o sistema de equações correspondente para as variáveis endógenas do modelo acima suposto (figura 5.3c) pode ser escrito como:

$$Y_1 = p_{Y_1X_3}X_3 + p_{Y_1X_5}X_5 + e_{Y_1}$$

$$Y_2 = p_{Y_2Y_1}Y_1 + p_{Y_2X_1}X_1 + e_{Y_2}$$

Os coeficientes para cada caminho (aresta ou arco) pode ser obtidos a partir de um programa computacional para uma regressão múltipla padrão. Também podem ser obtidos a partir dos dados na tabela 5.7. Cálculos feitos para todos os coeficientes são mostrados na figura 5.3d abaixo.

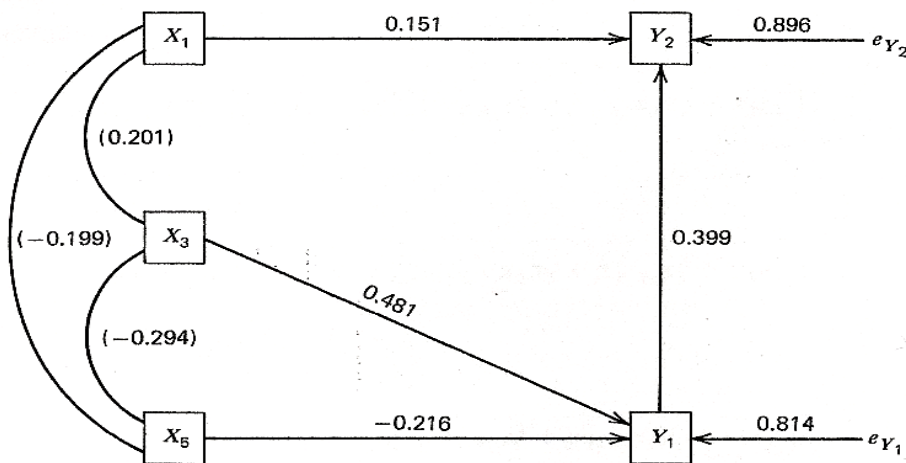


Figura 5.3d: grafo com os coeficientes dos efeitos entre as variáveis no modelo.

5.4. Análise de séries temporais

Uma *série temporal* é uma coleção de observações feitas sequencialmente ao longo do tempo. A característica mais importante deste tipo de dado é que as observações vizinhas são dependentes e estamos interessados em analisar e modelar esta dependência.

Enquanto em modelos de regressão, por exemplo, a ordem das observações é irrelevante para a análise, em séries temporais a ordem dos dados é crucial. Vale notar também que o tempo pode ser substituído por outra variável como espaço, profundidade, etc.

Como a maior parte dos procedimentos estatísticos foi desenvolvida para analisar observações independentes o estudo de séries temporais requer o uso de técnicas específicas.

Dados de séries temporais surgem em vários campos do conhecimento como Economia (preços diários de ações, taxa mensal de desemprego, produção industrial), Medicina (eletrocardiograma, eletro-encefalograma), Epidemiologia (número mensal de novos casos de meningite), Meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento), etc.

Algumas características são particulares a este tipo de dado, por exemplo,

- Observações correlacionadas são mais difíceis de analisar e requerem técnicas específicas.
- Precisamos levar em conta a *ordem temporal* das observações.
- Fatores complicadores como presença de tendências e variação sazonal ou cíclica pode ser difícil de estimar ou remover.
- A seleção de modelos pode ser bastante complicada, e as ferramentas podem ser de difícil interpretação.
- Devido à natureza sequencial é mais difícil de lidar com observações perdidas e dados discrepantes.

As séries temporais podem ser modeladas em forma de grafos, no qual o conjunto de vértices consistirá de componentes da série, enquanto as arestas representarão uma estrutura de correlação parcial dos componentes dado os outros componentes.

De acordo com Bach e Jordan (2003), os modelos de grafos para séries temporais são geralmente definidos no domínio de tempo. Isto é, eles definem uma distribuição de probabilidade de transição de um conjunto de variáveis de um estado, condicionado nos valores dessas variáveis no tempo das etapas anteriores.

Na prática, antes de se conduzir qualquer análise é importante definir se a série é estacionária ou não, para, a partir daí, estabelecer a estrutura do modelo probabilístico que estimará a série.

Uma série é considerada estacionária quando suas observações ocorrem ou oscilam aleatoriamente ao redor de uma média constante, ou seja, não há tendência. A condição de estacionariedade para uma série de segunda ordem implica em:

$$E(Y_t) = E(Y_{t+m}) = \mu_t \text{ (condição de média constante)}$$

$$Var(Y_t) = Var(Y_{t+m}) = \sigma^2 \text{ (condição de variância do processo é constante) e}$$

$$Cov(Y_t, Y_s) = C(t, s) = C(t - s) \text{ é função apenas de } t - s.$$

Tanto o modelo de Média Móvel em inglês *Moving average* (MA), quanto o modelo Auto-regressivo (AR), quanto o modelo Misto (ARMA) são utilizados para séries estacionárias.

Em modelo grafo clássico de série temporal estacionária com variáveis gaussianas ou com distribuição multivariada normal, a independência condicional pode ser mostrada através de valor zero na matriz de covariância inversa.

Assim, duas séries temporais y_i e y_j são condicionalmente independentes dado y_k se somente se $\{y_i(t), t \in \mathbb{Z}\}$ e $\{y_j(t), t \in \mathbb{Z}\}$ são condicionalmente independente dado $\{y_k(t), t \in \mathbb{Z}\}$.

Seja $y(t) = (y_1(t), \dots, y_3(t))$ uma série temporal multivariada, em que cada $t \in \mathbb{Z}$. A estrutura ou grafo do modelo para esta série temporal é apresentada na figura 5.4.

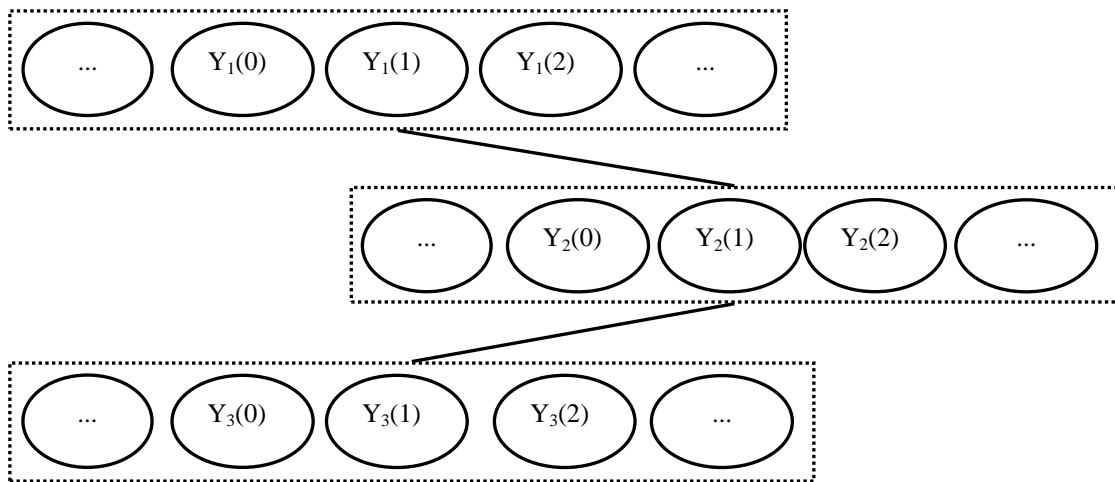


Figura 5.4: Modelo gráfico de três séries temporais.

Portanto, assumindo que as séries temporais na figura acima são estacionárias e gaussianas, temos que y_1 e y_3 são condicionalmente independentes dado y_2 .

5.4.1. Modelos de Média Móveis

Os modelos de média móvel utilizam como previsão para um determinado período no futuro a média das observações passadas. Portanto, Trata-se de um método que tem por objetivo “suavizar” as variações das séries por um processo de sucessivas médias.

Mediante o emprego de médias móveis, podem ser eliminadas as variações sazonais e irregulares, conservando-se desta forma o movimento de tendência. As médias móveis podem ser simples, centradas ou ponderadas.

O termo “média móveis” é utilizado porque à medida que a próxima observação está disponível, a média das observações é recalculada, incluindo esta observação no conjunto de observações e desprezando a observação mais antiga.

Seja e_t aleatório com média zero e variância $\sigma_{e_t}^2$, onde e_t é o termo que represente os erros aleatórios ou ruídos, ou ainda distúrbios. O modelo de médias móveis MA(q) assume que a série modelada é gerada através de uma combinação linear de q sinais de ruídos $e(t - i)$ aleatórios e independentes entre si,

Formalmente o modelo Y_t de médias móveis de ordem q , ou $MA(q)$ é representado por

$$Y_t = e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \dots + \beta_q e_{t-q}$$

A estrutura ou grafo para um modelo de média móvel de ordem três (MA (3)) é apresentada na figura abaixo (figura 5.4.1).

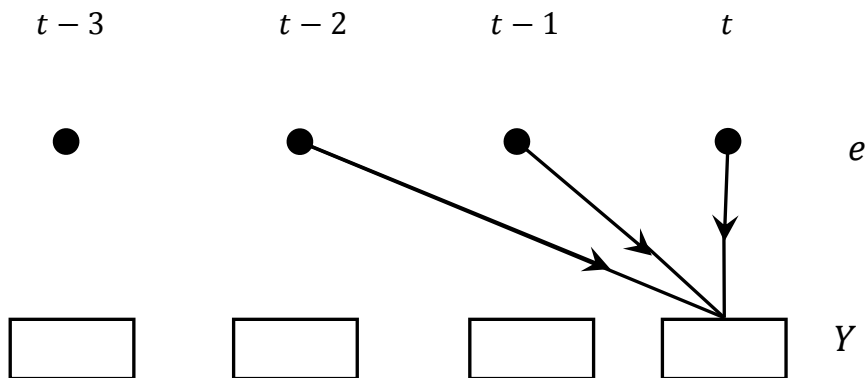


Figura 5.4.1: grafo para uma estrutura de um modelo de média móvel de ordem três.

5.4.2. Modelo Auto-regressivo

Os modelos Auto-regressivos (AR) são similares a um modelo de regressão múltipla, onde os valores passados de Y_t fazem o papel das regressoras.

O modelo AR de ordem p pode ser formalmente dado por:

$$Y_t = e_t + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p}$$

A figura abaixo (figura 5.4.2) ilustra uma estrutura de um modelo Auto-regressivo de ordem três.

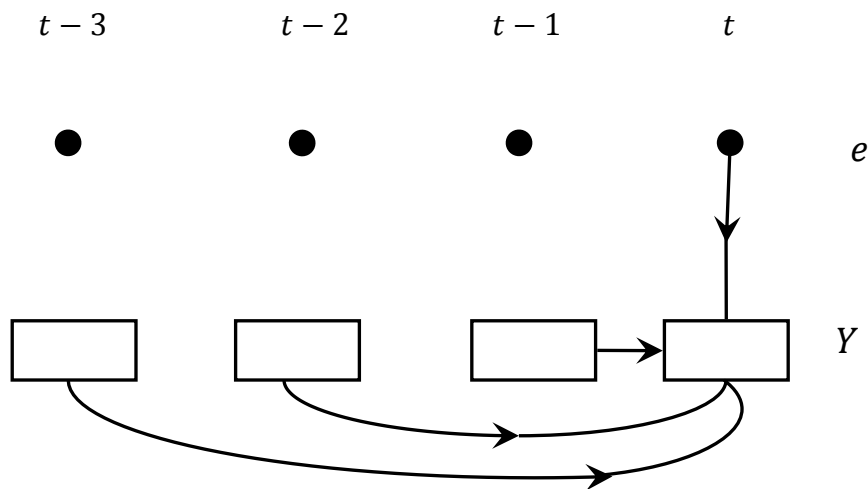


Figura 5.4.2: grafo para uma estrutura de modelo auto-regressivo de ordem três.

No modelo auto-regressivo de primeira ordem $AR(1)$ existe uma estrutura Markoviana se for razoável assumir que o valor atual de uma série temporal depende do seu passado imediato mais um erro aleatório. No sentido de que, dado um Y_{t-1} , o Y_t não depende de Y_{t-2}, Y_{t-3}, \dots , mas sim depende de Y_{t-1} mais um erro aleatório e_t .

5.4.3. Modelo Misto (Média Móvel e Auto-regressivo “ARMA”)

A combinação dos modelos AR(p) e MA(q) dá então origem ao modelo ARMA (p,q). Processos ARMA (auto-regressivo médias móvel) formam uma classe de modelos muito úteis e parcimoniosos para descrever dados de séries temporais.

O modelo ARMA de ordem (p, q) pode ser dados por,

$$Y_t = e_t + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \dots + \beta_q e_{t-q}$$

Onde p e q são os comprimentos dos *lags* (defasamento) dos termos auto-regressivos e termos de erro por médias móveis respectivamente.

Por exemplo, para uma situação que temos um modelo ARMA (1,2) teremos a seguinte estrutura de grafo (figura 5.4.3):

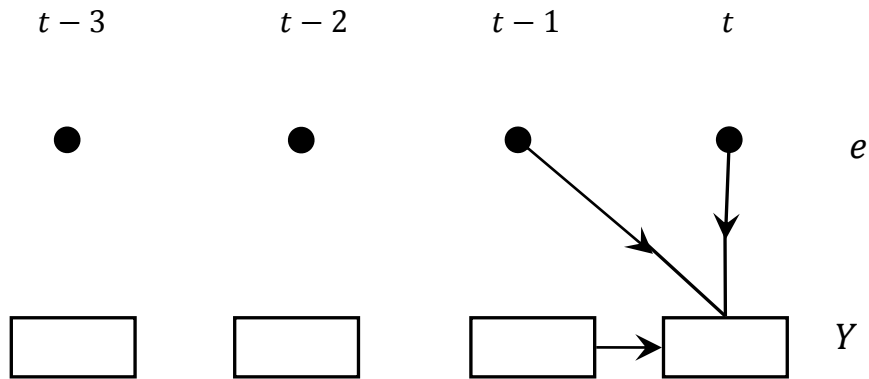


Figura 5.4.3: grafo para uma estrutura de modelo ARMA (1, 2).

5.4.4. Modelos de Função de Transferência

Em alguns casos a análise de séries temporal pode ser combinada com a técnica de regressão linear, com objetivo de gerar melhores previsões do que o uso isolado de um desses dois métodos. A partir desta combinação surge o termo Função Transferência, onde a idéia básica é adicionar informações que não fazem parte dos valores passados da própria série em estudo, através da inserção de variáveis explicativas, freqüentemente conhecidas por variáveis independentes.

Portanto, assume-se uma relação casual entre as variáveis dependente e independente que forneça as informações acima mencionadas. As figuras 5.4.4a e 5.4.4b ilustram alguns exemplos da estrutura ou grafo para modelos de variáveis defasadas (lagged variables model L(3)) e um modelo de função de transferência que é combinação de L(3) + ARMA (1, 2). Onde "X" é uma variável explicativa (independente) e u_t são uma seqüência independente de variáveis aleatórias.

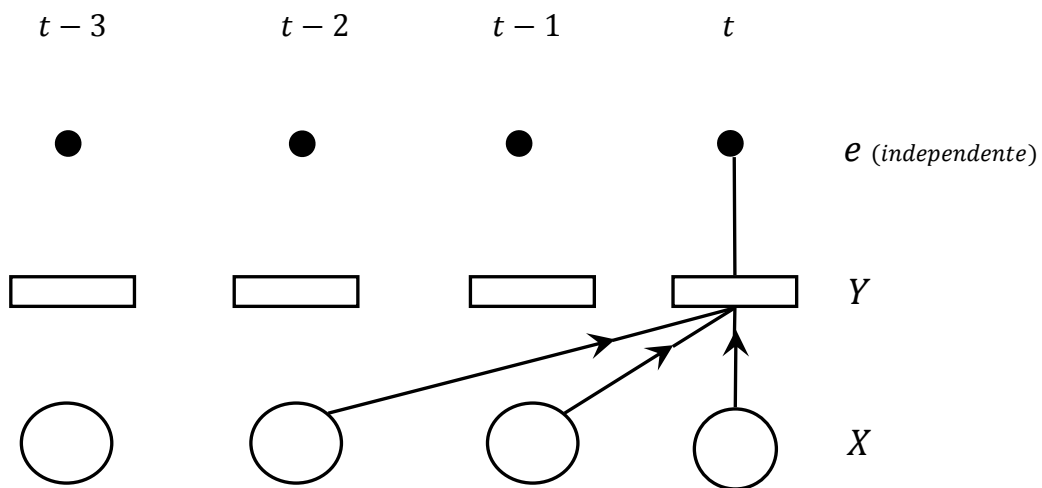


Figura 5.4.4a: grafo de um modelo de variáveis defasada L(3).

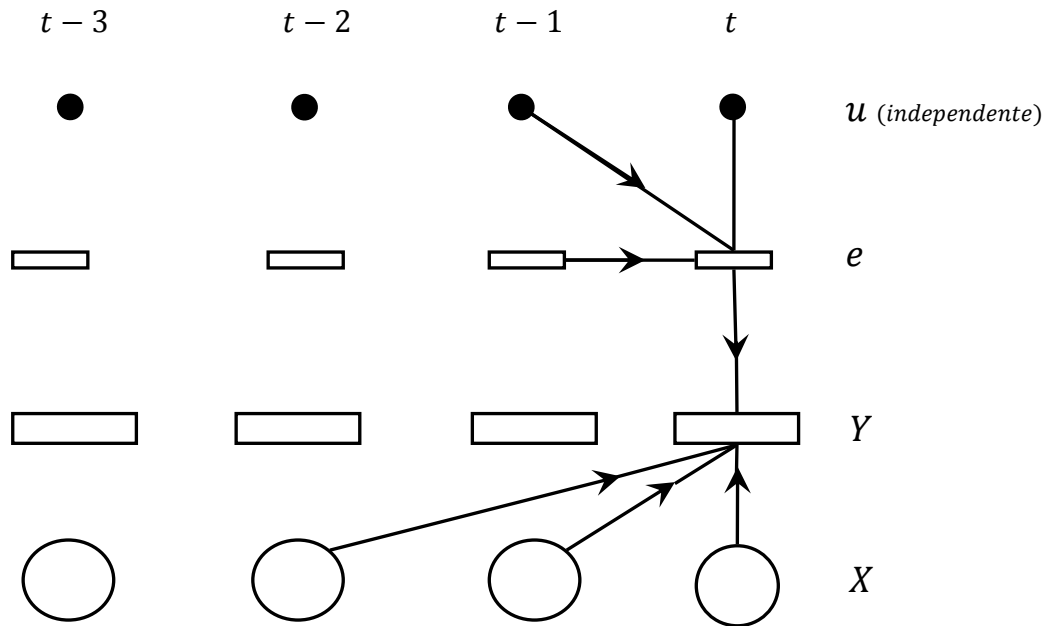


Figura 5.4.4b: grafo de um modelo de função de transferência L(3) + ARMA (1,

Uma extensão natural do modelo de função de transferência consiste em situações onde Y_t não depende apenas X_t e seus valores do passado (X_{t-1}, X_{t-2}, \dots) mas também depende dos seus próprios valores passados. Estes modelos são chamados de Modelo de Equação de Diferença Estocástica e denotados por $D(p, q)$.

A forma geral de tais modelos é,

$$Y_t = e_t + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} + \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \dots + \beta_q x_{t-q} + e_t$$

Este modelo é uma mistura entre modelo auto-regressivo e modelo de função de transferência, a essencial diferença deste modelo com o modelo de função de transferência é a introdução de feedback, no sentido de que mudanças em y_{t-1} , por exemplo, são diretamente reintroduzidas na equação para influência de y_t .

Estes modelos são apresentados estruturalmente de acordo com a figura 5.4.4c e 5.4.4d.

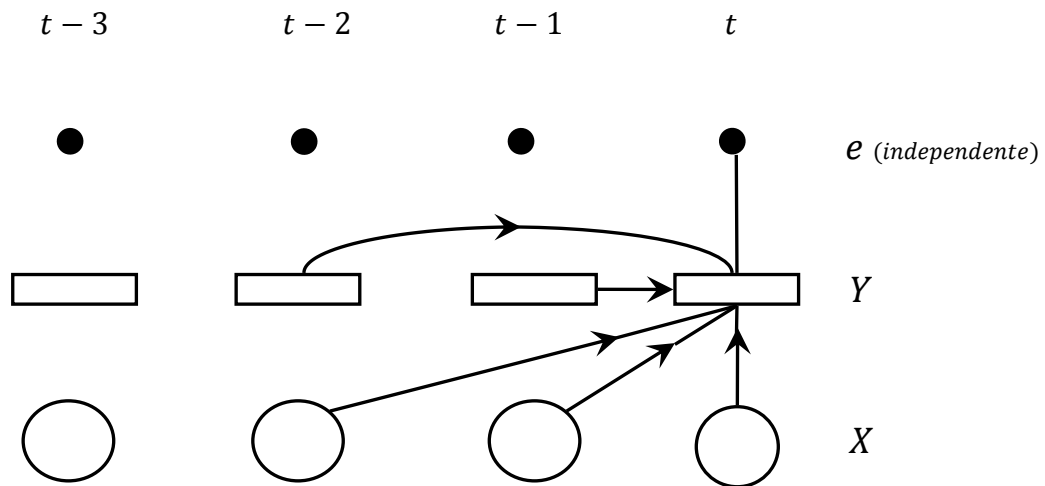


Figura 5.4.4c: grafo de um modelo de diferença estocástica $D(2, 3)$.

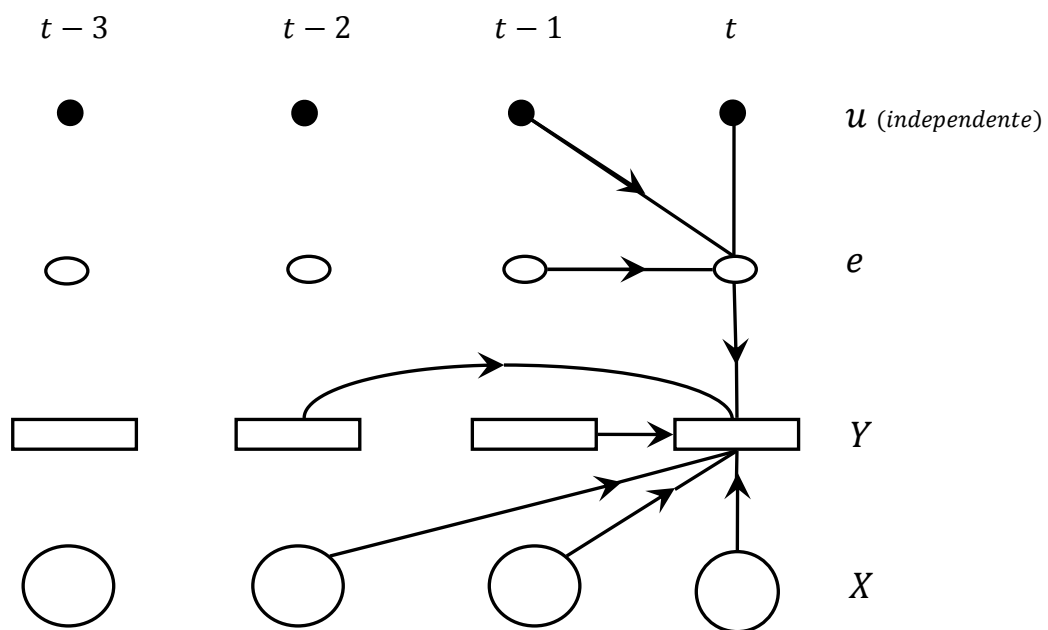


Figura 5.4.4d: grafo de um modelo de diferença estocástica $D(2, 3) + ARMA$

5.5. Modelos em espaço de estados

Os modelos em espaços de estados são modelos usados para previsão de séries temporais baseados na independência do futuro do processo em relação ao passado, dado o estado presente. Ou seja, toda informação do passado está contida no estado do processo, e isso é suficiente para a predição. Estes modelos são representados por duas equações dinâmicas:

$$S_t = AS_{t-1} + BX_t + \varepsilon_t$$

$$Y_t = CS_t + e_t$$

A primeira equação de transição (S_t), que representa a evolução dinâmica do vetor de estado não observado, e segunda equação de medida Y_t , representa as observações do processo geradas em função do vetor de estado (uma série de variável endógena). As duas perturbações (ε_t, e_t), não são correlacionadas entre si para qualquer defasagem, além de serem normais e auto-independentes e X_t representa uma série de variável exógena, Brillinger (1996).

Esse modelo espaço de estado pode ser usualmente descrito por um grafo (figura 5.5).

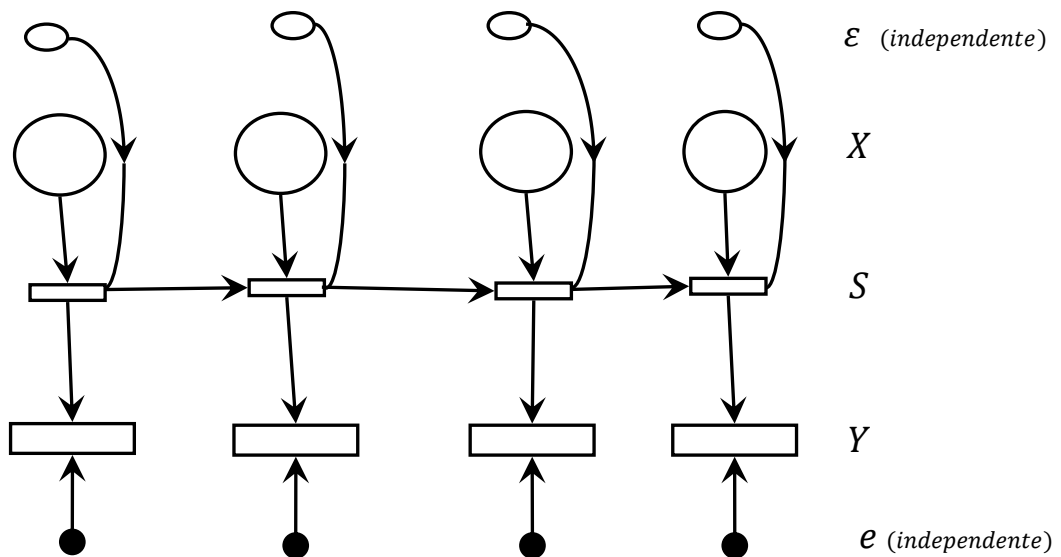


Figura 5.5: grafo do modelo espaço de estado com a variável exógena X_t , a variável estado S_t e a variável endógena Y_t .

Para se fazer uma previsão ou extrapolações de valores futuros através deste modelo na forma de espaço de estados, é preciso algo que produza estimadores atualizados do vetor de estado, que não é observado.

O algoritmo para esse fim é o chamado Filtro de Kalman, que consiste em quatro equações: duas equações de previsão (utilizando até o tempo $t-1$) e duas equações de atualização, permitindo "a atualização seqüencial do vetor de estado no instante $t-1$ para o instante t ", Kang, (2007) citando Souza, (1989, p. 61).

5.6. Modelos Mecânicos em Dinâmicas Econômicas

Modelos mecânicos em Dinâmicas Econômicas ou modelo hidráulico da economia, segundo Phillips (1950) foram desenvolvidos na Universidade de Leeds, numa tentativa de através de um modelo mecânico (uma maquina) permitir uma visualização de alterações quantitativas que ocorrem em sistemas de variáveis inter-relacionadas, seguido de mudanças iniciais em um ou mais deles.

Como um exemplo conceitual de uma rede econômica, de algum interesse histórico, pode-se mencionar a máquina Philips que foi concebida por Bill Phillips (1914-1975), um neozelandês engenheiro que virou economista. Phillips projetou a máquina para demonstrar de forma visual o fluxo circular do dinheiro dentro da economia.

Essa máquina mundialmente conhecida por MONIAC (**Monetary National Income Analogue Computer**) era de aproximadamente 2 m de altura, 1,2 m de largura e quase 1m de profundidade, e consistia em uma série de tanques e tubulações de plástico transparente que estava preso a uma tábua de madeira. Cada tanque representado algum aspecto da economia nacional do Reino Unido e do fluxo de dinheiro ao redor, a economia foi ilustrada por água colorida.

No topo do conselho há um grande tanque chamado de tesouraria. Água (que representa dinheiro) fluiu do tesouro para outros tanques que representam as diversas formas em que um país pode gastar seu dinheiro. O fluxo da água representa o movimento do dinheiro e a acumulação de dinheiro é representada pela captação de água nos tanques. Resultados são lidos a partir de escalas localizadas em vários lugares.

O fluxo real de água era controlado automaticamente através de uma série de flutuadores, contrapesos, eletrodos e cabos. Quando o nível da água atingia um determinado nível em um tanque, a bombas e drenos eram ativados. Para sua surpresa, Phillips e seu sócio Walter Newlyn, descobriram que a máquina poderia ser calibrada com uma precisão de $\pm 2\%$.

Numa das modelagens, envolvendo as variáveis como, consumo, despesa total, despesas de investimento, Saldo Positivo, Saldo de Transferência, Poupanças e Renda foi estruturado um grafo, em que as setas indicam o sentido do fluxo de líquidos na máquina.

Numa das modelagens, água foi bombeada para cima e os fluxos para baixo. Poupanças (S) são desviadas para um ciclo separado e caem em um tanque (M'). Partes destes voltam ao fluxo principal como despesa (E) e investimento (I).

O fluxo líquido se acumula no fundo do tanque. O nível do reservatório representa o saldo positivo (M), necessário para um determinado nível de atividade econômica. O líquido é bombeado de volta para o topo.

Este modelo de Phillips foi estruturado em forma de grafo, em que as setas indicam o sentido do fluxo de líquidos na máquina.

O fluxo é iniciado, por exemplo, alterando as despesas. As séries aqui envolvidas são temporais (contínua). A figura 5.6 mostra o grafo do modelo simples de Phillips.

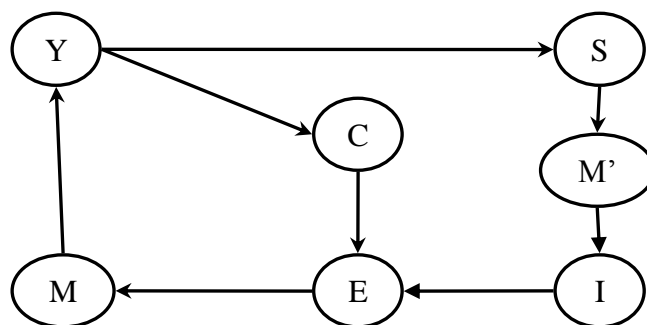


Figura 5.6: grafo representando um modelo simples de Phillips

C: consumo, E: Despesa total, I: Despesas de investimento, M: Saldo Positivo, M': Saldos de Transferência, S: Poupanças e Y: Renda.

Por exemplo, no modelo acima, quando o fluxo de poupança excede o fluxo de investimentos, o nível de água no tanque entre a poupança e investimento subiria para refletir o saldo acumulado. Quando o fluxo de investimento excede o fluxo de poupança para qualquer período de tempo, o tanque de saldos de transferência iria secar.

Capítulo 6

Considerações Finais

Nesta dissertação no seu capítulo segundo e terceiro apresentamos uma introdução ou fundamentos da teoria de grafos onde nele foram apresentadas varias definições de alguns conceitos relacionados com a teoria de grafos.

No capítulo quarto foram apresentados alguns conceitos probabilísticos, concretamente a independência condicional através da propriedade de Markov, mas adiante, no capítulo seguinte foram descritos os conceito de redes Bayesianas, diagramas de influencias que através de alguns exemplos foi ilustrado a sua utilidade pratica para modelagem de alguns problemas do cotidiano.

A modelação com grafos em modelos Multivariados foi apresentada no capítulo sexto, onde, por exemplo, foram ilustrados exemplos Multivariados para variáveis discretas (modelo log-linear) e para variáveis contínuas (Modelos Gaussianos).

Os exemplos foram baseados em Whittaker (1990) de aplicação metodologia de um conjunto de dados, mostrando a relação de notas em cinco disciplinas, isto no caso gaussiano, e para o caso discreto (log-linear) com base em Edwards (1995) onde foi ilustrado um modelo para tabela de contingências de multi-entradas, baseadas em amostragem da distribuição Multinomial.

Também foi apresentada a independência condicional entre variáveis em modelo grafo gaussianos ou de variáveis com distribuição multivariada normal, onde através da matriz inversa de covariâncias ou da matriz inversa escalonada de coeficientes correlações é possível distinguir as relações entre variáveis.

Sendo que um valor zero em uma célula da matriz acima citado, indica de certo modo uma independência condicional entre as variáveis correspondentes, o que implica uma ausência de aresta entre as variáveis em causa na estrutura do grafo.

Mas adiante foram apresentados os modelos de regressão linear simples, múltipla, regressão multivariada, equações estruturais, e serie temporais. Em que neles são ilustrados alguns grafos para descrever a estrutura gráfica para cada modelo e cada um descritos com base na sua utilidade para modelagem e resolução de muitos problemas em estatística.

Um grafo foi apresentado, resultante de um estudo duma simulação através de uma máquina o comportamento da economia Britânica nos anos quarentas, feito pela universidade de Leeds, numa tentativa de visualizar as alterações quantitativas que ocorrem em sistemas de variáveis inter-relacionadas, seguido de umas alterações iniciais em algumas variáveis no sistema. Essa máquina (denominado por Máquina de Philips) ganhou um interesse histórico e o modelo resultante é conhecido por Modelo mecânico da dinâmica econômica.

Espera-se que o conteúdo apresentado nesta dissertação ajude de certa maneira a todas as pessoas que estejam interessadas na modelagem de vários problemas probabilísticos e de modelos estatísticos multivariados através de grafo.

Para trabalhos futuros, sugerimos uma melhoria de alguns aspectos que provavelmente não foram claramente aqui apresentados, e que, ao em vez de se limitar apenas na modelagem, sejam incluídos alguns algoritmos nos grafos que permitam resolver duma maneira eficiente os diversos problemas e relacionar os algoritmos nos grafos com os algoritmos de estatística ou probabilísticos para os problemas aqui citados.

Referencias Bibliográfica

1. Bach, F. R. e Jordan, M. I. (2003). *Learning Graphical Models for Stationary Time Series*. Technical Report 650. Department of Statistics, University of California, Berkeley.
2. Bhansali, R.J. e Ippoliti, L. (2005). “Inverse Correlations for Multiple Time Series and Gaussian Random Fields and Measures of Their Linear Determinism”, *Journal of Mathematics and Statistics* v.1, n. 4, pp. 287-299;
3. Boaventura Netto, P. O. (1996). *Grafos: Teoria, Modelos, Algoritmos*, 2a. ed, Edgard Blucher;
4. Boaventura Netto, P. O. Jurkiewicz, S (2009). *Grafos: Introdução e Práticas*, Edgard Blucher LTDA;
5. Brillinger, D. R. (1996). “Remarks Concerning Graphical models for Time Series and Point Processes”, *Revista Brasileira de Economia* v.16, pp. 1-23;
6. Christensen, R. (1997). *Log Linear Models and Logistics Regression*, Second Edition, Springer texts in statistic;
7. Cox, D. R., Wermuth, N. (1996). *Multivariate Dependencies Models, Analysis and Interpretation*. Chapman & Hall/CRC;
8. Cox, D.R. & Wermuth, N. (2004). “Causality: a statistical view”, *International Statistical Review*, v. 72, n. 3, pp. 285–305, Printed in Wales by Cambrian Printers, International Statistical Institute;
9. Dillon, W. R. G. M. (1984). *Multivariate Analysis Methods and Applications*. John Wiley;
10. Edwards, D. (1995). *Introduction to Graphical Modelling*. 2ª ed, Springer-Verlag New York Inc;

11. Eichler, M. (2006). *Fitting Graphical Interaction Models to Multivariate Time Series*. Department of Quantitative Economics University of Maastricht, Netherlands. Disponível em: <http://www.personeel.unimaas.nl/m.eichler/pub.html> acessado em 05/10/2009;
12. FARIAS, S.A.; SANTOS, R.C. (2000). “Modelagem de Equações Estruturais e Satisfação do Consumidor: uma Investigação Teórica e Prática”. *Revista de Administração Contemporânea - RAC*, v.4, n.3, pp 107-132;
13. Gilechrist, W. (1984). *Statistical Modelling*. John Wiley e Sons;
14. Hair, J.; Tatham, R. L.; Anderson, R. E. e Black, W. (1998). *Análise Multivariada de Dados*. 5ª edição, Laser House;
15. Hayes, B. (2000). “GRAPH THEORY IN PRACTICE: PART I”. *The American Scientific Research Society*, v. 88, pp. 9-13;
16. Jensen, F.V. (1996). *An Introduction to Bayesian Networks*, Springer-Verlag, New York;
17. Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs*, Springer;
18. Lauritzen, S. L. and Wermuth, N. (1989). “On substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models”. *Journal of the Royal Statistical Society*, v. 52, pp. 21-50;
19. Lauritzen, S.L. (2006). *Fundamental Elements of Graphical Models*. Springer;
20. M. Eichler and V. Didelez (2007). “Causal reasoning in graphical time series models”. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*. Disponível em <http://www.personeel.unimaas.nl/m.eichler/pub.html>, acesso em 12/01/2010;

21. Marchetti, G. M. (2006). “Independencies Induced from a Graphical Markov Model After Marginalization and Conditioning: The R Package ggm”. *Journal of Statistical Software* – volume 15, issue 6.
22. Nogueira, F. E. (2007). *Modelos de Regressão Multivariada*. Tese de Mestrados no Instituto de Matemática da USP. Disponível em: “<http://www.teses.usp.br/teses/disponiveis/45/45133/tde-25062007-163150/>”, acessado em 09/09/2009, 12h25;
23. Oliveira, R.D. e CARDOSO, M.R.A. (2001). “Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos”. *Revista Brasileira de Epidemiologia*, v. 4, pp. 145-152;
24. Phillips, A. W. (1950). “Mechanical model in economic dynamics”. *Economica, New Series*, v. 17, pp. 283-305;
25. Rocha, C., Santana, Á. L., Francês, C. R., Rego, L., Costa, J., Carvalho, S. V. e Vijaykumar, N. L.. (2006). *Aplicação de Modelos Markovianos para a Análise Temporal e Melhoria da Interpretabilidade de Redes Bayesianas*. Trabalho de pesquisa na Universidade Federal de Pará. Disponível em: http://www.lac.inpe.br/~vijay/download/Papers/Sbpo2007_1.pdf acessado em 05/10/2009, 21h34;
26. WASSERMAN, L (2004). *All of statistics: a concise course in statistical inference*. Springer texts in statistics. USA;
27. Whittacker, J. (1990). *Graphical Models in applied multivariate statistics*. John Wiley e Sons;
28. Wright, S. (1934). “The Method of Path Coefficients”: *The Annals of Mathematical Statistics*, v. 5, n. 3, pp. 161-215. Disponível em: <http://www.jstor.org/stable/2957502>, acessado em 16/02/2010, 10h41.