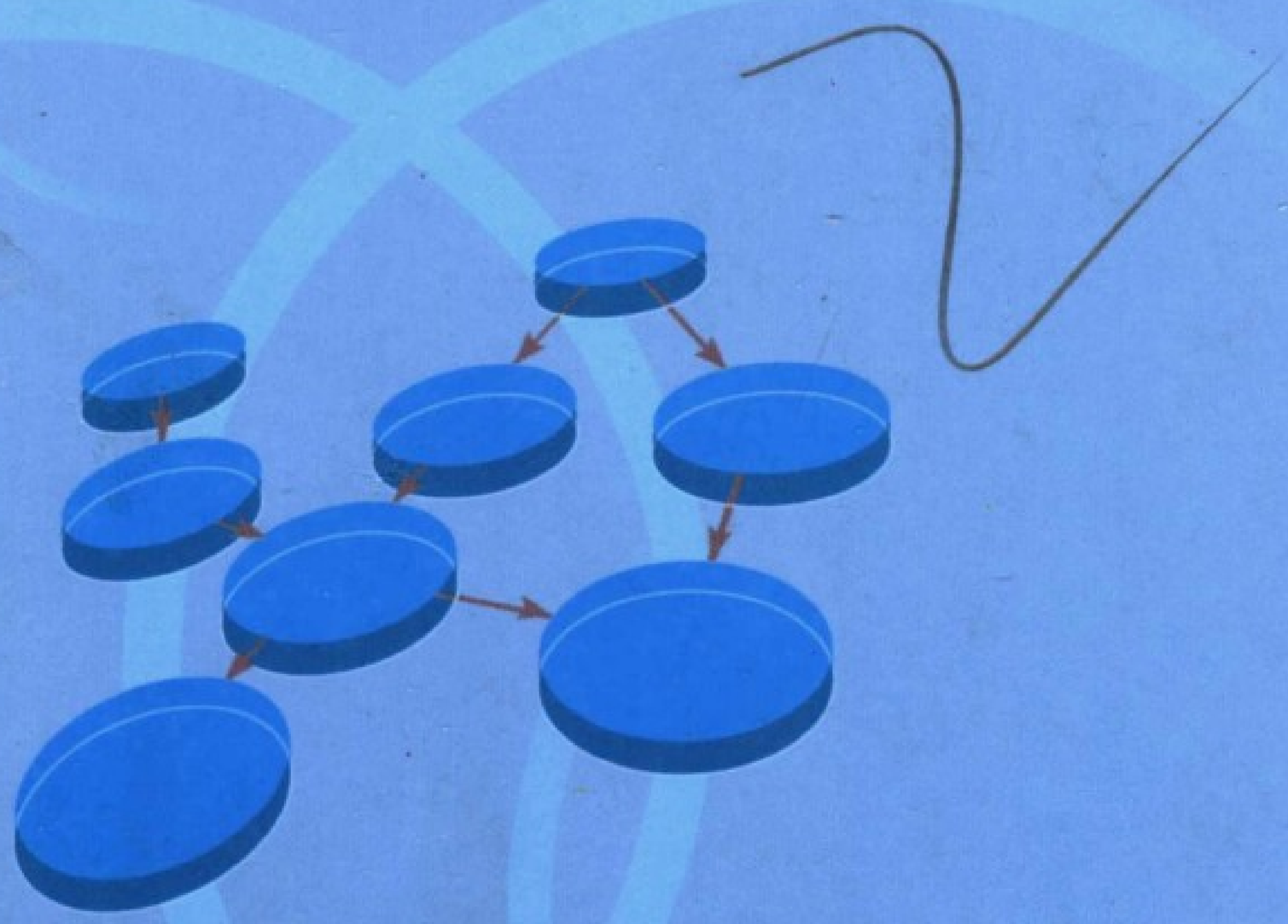
INTRODUCTION TO
BAYESIAN NETWORKS

贝叶斯网引论

张连文
郭海鹏 著

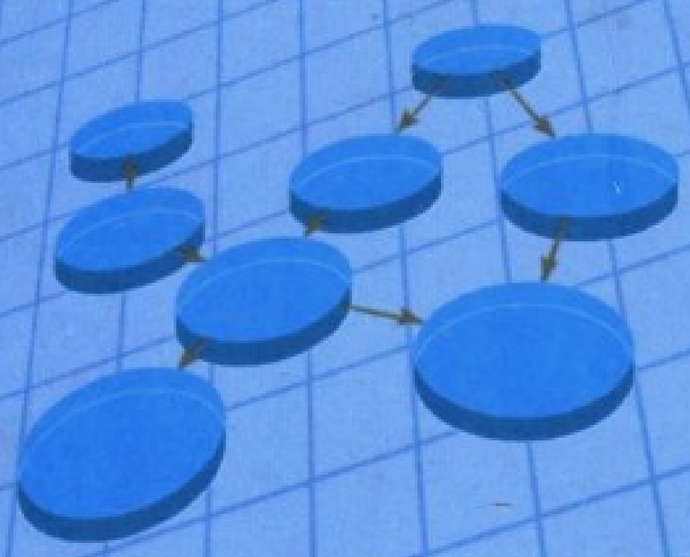


科学出版社
www.sciencep.com

(O-2610.0101)

INTRODUCTION TO BAYESIAN NETWORKS

贝叶斯网引论



ISBN 7-03-018170-0

9 787030 181701 >

ISBN 7-03-018170-0

定 价：58.00 元

O212
61
2006

中国科学院科学出版基金资助出版

贝叶斯网引论

张连文　郭海鹏　著

科学出版社
北　京

内 容 简 介

贝叶斯网是将概率、统计应用于复杂系统的不确定性推理和数据分析的一种有效工具，它起源于20世纪80年代中期对人工智能中的不确定性问题的研究，近年来在国际上的影响不断扩大。本书是第一本系统论述贝叶斯网的基本理论、算法及其应用的中文专著。内容包括概率论及贝叶斯网基本概念、贝叶斯网推理、贝叶斯网学习，以及贝叶斯网在中医中的应用四大部分。本书从实例出发，由浅入深，直观与严谨相结合，并提供了详尽的参考文献。本书的读者对象是相关专业的高年级本科生、研究生和科研人员。

图书在版编目(CIP)数据

贝叶斯网引论/张连文，郭海鹏著．—北京：科学出版社，2006
ISBN 7-03-018170-0

Ⅰ．贝…　Ⅱ．①张…②郭…　Ⅲ．贝叶斯推断　Ⅳ．O212

中国版本图书馆CIP数据核字(2006)第125753号

责任编辑：赵卫江　陈砺川/责任校对：赵　燕
责任印制：吕春珉/封面设计：耕者设计工作室

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com
双青印刷厂 印刷
科学出版社发行　各地新华书店经销
*
2006年11月第　一　版　　开本：B5(720×1000)
2006年11月第一次印刷　　印张：19
印数：1—2 000　　字数：360 000

定价：58.00元

（如有印装质量问题，我社负责调换〈新欣〉）
销售部电话 010-62136131　编辑部电话 010-62138017(BI01)

序

这是一本系统论述由概率、统计与图论结合而发展起来的贝叶斯网的专著. 贝叶斯网起源于 20 世纪 80 年代中期对人工智能中的不确定性问题的研究，已经成为人工智能的一个重要领域. 近年来它的影响在国际上不断扩大，成为将概率、统计应用于复杂系统的不确定性推理和数据分析的一种有效工具. 本书作者根据多年来在这方面的研究和与国际交流的体会，采取由实例出发、由浅入深、直观与严谨相结合等较易为读者理解的方式，对这一新的学科的基本理论、方法及其应用进行了系统的、较全面的介绍，列举了相当详尽的参考文献. 我认为，本书在国内的出版是一桩在学术上有意义的事件，将促进国内对这一新兴而又有潜能的领域的进一步学习、研究和应用. 作者希望我给本书写序，作为人工智能研究的一个热心观众，借这个机会谈一点有关人工智能以及它与数学、电子计算机（下面采用现在通用的称谓——电脑）的关系的粗浅看法，供有志于人工智能研究的读者参考.

人工智能的兴起大致源于 20 世纪 40 年代电脑的出现. 人们一方面兴奋地运用电脑强大的计算功能，去完成以往无法完成以及新提出的庞大计算任务；另一方面，对电脑所具有的类似于人脑的某些功能，包括一些逻辑判断功能，产生了巨大兴趣，从而对它寄予了很高的期望，期望它能够在智能方面起作用. 于是人们开始研究自动机的理论、人类智能的机理以及智能行为的计算模型，进而期望构造与人脑相类似的智能系统，这大概是人工智能的最初想法. 在当时，已经有人认为要制造完全代替人脑的电脑是不可能的. 50 年代，我从著名的数理逻辑界前辈胡世华先生那里知道了一种相当有说服力的说法（我认为现在看来也是正确的）：机械是人手的延伸，而电脑是人脑的延伸. 机械常常能够更好、更快地完成人手所能完成的机械性任务，但是不能完全代替人手；同样地，电脑能够完成人脑所能完成的机械性任务，但是不可能完全代替人脑. 当然这种看法丝毫没有贬低人工智能的意义，相反，人工智能是一个引起人们浓厚兴趣、已经有相当成就而又被人们寄予巨大希望的科学领域.

与探索和模拟人类智能这一长远目标相比，人工智能研究更多的是集中在针对实际具体问题的智能系统的研究和开发. 专家系统就是一个例子，它的目标是将专家们对某一复杂领域的知识和经验引入电脑系统，使得一般人能够借助电脑系统得到对问题的专家级解答. 60 年代，在国内由于肝炎的流行，人们研制出学习和模仿著名中医专家关幼波的诊治系统就是一例. 正如本书的 1.4 节所表述

的，人工智能的实质进展有赖于不断针对人类的某种智能行为，运用数学理论和方法，结合计算机技术来建立适当的数学计算模型. 只有智能的目标和计算机技术而没有数学的深层次介入是不可能有显著进展的. 不确定性是人工智能所面临的一个重要课题，因此概率论自然就成为处理人工智能问题的一个工具. 对于静态的概率问题的研究和计算，理论上只需要一个联合概率分布就足够了. 但是联合概率分布的复杂度相对于变量个数呈指数增长，以致当变量个数很大时，计算复杂度极高而不可行. 虽有电脑的强大计算功能也会成为不可行. 贝叶斯网的提出就是要解决这个问题. 它用图论的语言直观揭示问题的结构，又按照概率论的原则对问题的结构加以分析，降低推理计算的复杂度，使得人们能够将概率论应用于大型问题. 近年来贝叶斯网能够在众多领域得到广泛应用，原因就在于此. 从贝叶斯网的发展，我们同样也看到计算机技术的广泛应用有赖于数学的深层次介入.

对于概率论与数理统计的广泛应用，人们有普遍的共识. 但是在很长一段时期，人们普遍地关注和运用概率的频率解释，对于贝叶斯观点比较忽视，认为将概率看作一种主观信任程度是没有客观标准的. 20 世纪中、后期以来，贝叶斯观点的应用逐步得到了认同，贝叶斯方法再度兴起，贝叶斯网就与这次潮流密切相关，这些是概率论与数理统计的应用中一个值得重视的趋势. 现在看来，主观、客观观念属于哲学范畴，应用于具体的对象时，需要根据实践来判断. 主观信度或在贝叶斯观点下的先验概率，实际上可以认为是在本次试验或观察以前或以外的某种经验积累的表述. 所以，对于一种数学观点和方法，应该依据实践来认识、判断和应用，这样才不至于因噎废食.

本书作者之一张连文 1986 年在北京师范大学汪培庄教授的指导下获得应用数学硕士学位，接着在我们的概率统计学科点攻读博士，作为我和 Kansas 大学的 Glenn Shafer 教授联合指导的博士生，以论文《人工智能中不确定性的三个模型》通过论文答辩（1989）. 我当时觉得他比较年轻，对人工智能有兴趣，就建议他再在国外攻读一个计算机博士学位，这样对以后研究人工智能会有好处. 1994 年，他在加拿大不列颠哥伦比亚大学 David Poole 教授的指导下获得计算机博士学位，论文标题是《决策网络的计算理论》. 这些年来，他实际上一直围绕人工智能中不确定性问题而工作，对贝叶斯网的发展做出了一系列的重要工作；而且，他还和中医学者合作，应用贝叶斯网研究中医的辨证施治. 这次能在科学出版社出版《贝叶斯网引论》，是向内地对他的培养的一项汇报和宝贵的回报. 我希望通过本书的出版，使他能够和内地建立更好的学术交流，帮助推动贝叶斯网及相关方面在内地的应用和发展.

严士健
于北京师范大学数学科学学院

前　言

贝叶斯网（Bayesian networks）是一种帮助人们将概率统计应用于复杂领域、进行不确定性推理和数据分析的工具. 它起源于人工智能领域的研究，近年来对众多其它领域也产生了重要影响. 本书系统地介绍贝叶斯网的基本理论和方法，为读者了解和进入这个新兴领域提供一条相对平坦的途径.

从技术层面上讲，贝叶斯网是一种系统地描述随机变量之间关系的语言. 构造贝叶斯网的主要目的是进行概率推理，即计算一些事件发生的概率. 要在一些随机变量之间进行概率推理，理论上只需要一个联合概率分布即可. 但是，联合概率分布的复杂度相对于变量个数成指数增长，所以当变量众多时不可行. 贝叶斯网的提出就是要解决这个问题，它把复杂的联合概率分布分解成一系列相对简单的模块，从而大大降低了知识获取的难度和概率推理的复杂度，使得人们可以把概率论应用于大型问题.

贝叶斯网是概率论与图论相结合的产物，它一方面用图论的语言直观揭示问题的结构，另一方面又按照概率论的原则对问题的结构加以利用，降低推理的计算复杂度. 近年来，贝叶斯网之所以能够在众多不同领域得到广泛应用，其根本原因就在于此. 另外，由于贝叶斯网直观易懂，它也是学者们喜爱的讨论和交流工具.

统计学、系统工程、信息论以及模式识别等学科中许多经典的多元概率模型都是贝叶斯网的特例，包括朴素贝叶斯模型（naïve Bayes models）（Titterington et al.，1981）、隐类模型（latent class models）（Lazarsfeld and Henry，1968；Goodman，1974）、混合模型（mixture models）（Everitt and Hand，1981；McLachlan and Basford，1988）、隐马尔可夫模型（hidden Markov models）（Baum and Petrie，1966）、卡尔曼滤波器（Kalman filters）（Kalman，1960）等等. 贝叶斯网为这些模型提供了一个共同的框架，使得在一个领域获得的结果可以推广到其它领域. 更重要的是，它也为发展新模型提供了一个自然的框架. 例如，动态贝叶斯网（dynamic Bayesian networks）（Dean and Kanazawa，1990；Russell and Norvig，2003）就是最近几年在这个框架下发展起来的一类新模型，它主要用于对多维离散时间序列的监控和预测. 另一个例子是多层隐类模型（hierarchical latent class models）（见第 9 章），它是对隐类模型的推广，能够揭示观测变量（observed variables）背后的隐结构.

构造贝叶斯网的方法因问题而定. 有时，网络结构和参数由应用问题的定义

所决定. 例如，在解码问题中（见 2.6.2 节），网络结构完全取决于编码器的设计，而网络参数则由编码器的设计和信道的特征所决定. 在专家系统中，贝叶斯网的结构和参数往往通过咨询专家来获得. 近年来，贝叶斯网越来越多地被用于数据分析，成了数据分析的工具. 统计学视之为图模型的一种，而人工智能学科把从数据出发获得贝叶斯网的过程视为是机器学习的一个特例，称为贝叶斯网学习.

在国际上，贝叶斯网已经成为人工智能和机器学习教材的重要内容，同时还有多本专著已经或即将出版，这些均凸显贝叶斯网受到的重视. 我们从事贝叶斯网研究多年，觉得有责任把它系统地介绍到国内，于是决定写这本书.

本书分四个部分. 第一部分介绍贝叶斯网基础，包括第 1、2、3 章. 第 1 章回顾一些概率论及信息论中与贝叶斯网密切相关的概念和结果. 第 2 章以降低概率推理的复杂度为出发点，引入贝叶斯网的概念，介绍如何手工构造贝叶斯网，并给出诸多应用实例. 第 3 章讨论贝叶斯网的图论侧面与它的概率论侧面之间的密切关系.

第二部分介绍贝叶斯网推理，包括第 4、5、6 章. 第 4 章揭示贝叶斯网推理的原理，并给出最基本的推理算法，即变量消元算法. 第 5 章介绍另一个推理算法，即团树传播算法. 与变量消元法相比，团树传播法的主要优点是它使得两次不同推理的中间结果可以共享. 因此，当需要做多次推理时，团树传播法比变量消元法更为合适. 第 6 章介绍近似推理算法，包括随机抽样法和变分法以及最新的研究进展.

第三部分讨论贝叶斯网学习，包括第 7、8、9 章. 第 7 章假设已知网络结构，讨论参数估计方法. 第 8 章探讨在不知道网络结构的情况下，如何通过数据分析同时确定贝叶斯网的结构和参数. 第 9 章假设观测变量背后有一个隐结构，研究如何通过数据分析对这个隐结构进行推测.

第四部分只有一章，即第 10 章，它介绍的是贝叶斯网在中医研究中的一个应用实例. 贝叶斯网的应用有很多，之所以选用中医的例子是因为贝叶斯网有机会对中医现代化起重要推动作用，而中医也为贝叶斯网研究提出了一些有趣的新问题.

本书的读者对象是相关专业的高年级本科生、研究生和科研人员. 我们选择贝叶斯网最核心的内容，基于我们自己的体会将其组织起来，因此与国际上现有的专著相比，本书更适合作为贝叶斯网的入门教材. 当然，本书也可以用于自学，书中不仅详细阐述贝叶斯网的基本理论，还简要介绍它的各种应用实例，综述最新研究进展，并提供大量的参考文献以便读者进一步深入学习. 在注意理论深度和严谨性的同时，我们也注重理论背后的直观含意以及如何将理论应用于实际.

本书的筹划始于 2002 年秋. 那时，张连文在香港科技大学开设了一门关于贝叶斯网的研究生课程，备课时已为将来把讲稿扩充成书做了许多考虑. 具体写作开始于 2003 年秋，初稿完成于 2005 年夏. 陈弢和王燚几次仔细阅读初稿，提出了许多改进意见，并参与了一些章节的写作，我们表示衷心感谢. 严士健教授是张连文的博士导师，长期关心和支持张连文的工作，这次又拨冗为本书写序，在此特表感激之意. 本书的写作得到香港研究资助局项目 622105、国家重点基础研究发展计划（973 计划）No. 2003CB517101，以及香港科技大学博士后基金的资助，本书的出版得到中国科学院科学出版基金的资助，史忠植教授、刘际明教授提供了协助，陈砺川、赵卫江女士做了大量细致的编辑工作，这里一并致谢.

由于我们水平有限，书中难免有疏漏和错误，请读者不吝指正. 错误更正会在本书的网页（http://www.cs.ust.hk/bnbook/）上公布. 在此网页上，读者还可以找到一些关于贝叶斯网的信息，如相关学术组织、学术会议、学术期刊、学术团队以及常用软件等.

张连文　郭海鹏

2006 年春于香港清水湾

目　录

第一部分　贝叶斯网基础

第 1 章　概率论基础 …… 3

1.1　随机事件与随机变量 …… 3

1.2　概率的解释 …… 6

1.2.1　古典解释 …… 6

1.2.2　频率解释 …… 6

1.2.3　主观解释 …… 7

1.2.4　特性解释与逻辑解释 …… 9

1.3　多元概率分布 …… 10

1.3.1　联合概率分布 …… 10

1.3.2　边缘概率分布 …… 11

1.3.3　条件概率分布 …… 13

1.3.4　边缘独立与条件独立 …… 16

1.3.5　贝叶斯定理 …… 18

1.4　概率论与人工智能 …… 20

1.5　信息论基础 …… 21

1.5.1　Jensen 不等式 …… 22

1.5.2　熵 …… 24

1.5.3　联合熵、条件熵和互信息 …… 25

1.5.4　相对熵 …… 28

1.5.5　互信息与变量独立 …… 29

第 2 章　贝叶斯网 …… 31

2.1　不确定性推理与联合概率分布 …… 31

2.2　条件独立与联合分布的分解 …… 33

2.3　贝叶斯网的概念 …… 34

2.4　贝叶斯网的构造 …… 36

2.4.1　确定网络结构 …… 36

2.4.2 因果关系与贝叶斯网 …… 39
2.4.3 确定网络参数 …… 41
2.5 贝叶斯网的应用 …… 45
2.5.1 医疗诊断 …… 45
2.5.2 工业应用 …… 48
2.5.3 金融分析 …… 48
2.5.4 计算机系统 …… 49
2.5.5 军事应用 …… 50
2.5.6 生态学 …… 51
2.5.7 农牧业 …… 53
2.6 贝叶斯网对其它领域的影响 …… 53
2.6.1 生物信息学 …… 53
2.6.2 编码学 …… 57
2.6.3 机器学习 …… 61
2.6.4 时序数据和动态模型 …… 62
2.7 文献介绍 …… 65
第3章 图分隔与变量独立 …… 66
3.1 直观分析 …… 66
3.1.1 基本情况 …… 66
3.1.2 一般情况 …… 68
3.2 有向分隔与条件独立 …… 70
3.2.1 几个引理 …… 70
3.2.2 马尔可夫性 …… 71
3.3 有向分隔与无向分隔 …… 74
3.4 有向无圈图与联合概率分布 …… 77
3.5 文献介绍 …… 78

第二部分 贝叶斯网推理

第4章 贝叶斯网与概率推理 …… 81
4.1 推理问题 …… 81
4.1.1 后验概率问题 …… 81
4.1.2 最大后验假设问题 …… 82
4.1.3 最大可能解释问题 …… 83

4.2 变量消元算法 …… 83
4.2.1 概率分布的分解与推理复杂度 …… 83
4.2.2 消元运算 …… 85
4.2.3 算法描述 …… 86
4.2.4 一个例子 …… 88
4.3 复杂度分析 …… 89
4.3.1 复杂性的度量 …… 89
4.3.2 复杂度的计算 …… 90
4.4 消元顺序 …… 94
4.4.1 最大势搜索 …… 95
4.4.2 最小缺边搜索 …… 95
4.5 推理问题简化 …… 97
4.6 MAP 假设问题 …… 99
4.6.1 两个运算 …… 100
4.6.2 分解与计算复杂度 …… 102
4.6.3 变量消元 MAP 算法 …… 102
4.7 文献介绍 …… 105
第 5 章 团树传播算法 …… 106
5.1 团树 …… 106
5.2 一个变量后验概率的计算 …… 107
5.3 团树传播的正确性 …… 111
5.4 团树传播与计算共享 …… 113
5.5 每个变量的后验概率的计算 …… 115
5.6 团树的构造 …… 118
5.6.1 图消元与团树 …… 118
5.6.2 图消元构造团树算法的正确性 …… 121
5.6.3 极小团树 …… 122
5.6.4 t-团与 g-团 …… 123
5.7 文献介绍 …… 124
第 6 章 近似推理 …… 125
6.1 随机抽样算法 …… 125
6.1.1 重要性抽样法 …… 125
6.1.2 MCMC 抽样 …… 131
6.2 变分法 …… 133

6.2.1 朴素平均场法 …… 134
6.2.2 循环传播算法 …… 136
6.3 其它近似推理算法 …… 138
6.4 文献介绍 …… 138

第三部分 贝叶斯网学习

第7章 参数学习 …… 143
7.1 贝叶斯网与数据分析 …… 143
7.2 单参数最大似然估计 …… 144
7.3 单参数贝叶斯估计 …… 146
7.4 单变量网络参数估计 …… 150
7.5 一般网络最大似然估计 …… 151
7.5.1 最大似然估计的计算 …… 152
7.5.2 最大似然估计的性质 …… 155
7.6 一般网络贝叶斯估计 …… 157
7.7 缺值数据最大似然估计 …… 160
7.7.1 EM算法的基本思想 …… 161
7.7.2 EM算法的基本理论 …… 163
7.7.3 EM算法 …… 165
7.7.4 EM算法的收敛性 …… 166
7.8 缺值数据贝叶斯估计 …… 168
7.9 文献介绍 …… 171
第8章 结构学习 …… 172
8.1 似然函数与模型选择 …… 172
8.2 贝叶斯模型选择 …… 175
8.3 大样本模型选择 …… 178
8.4 其它模型选择标准 …… 180
8.5 模型优化 …… 182
8.5.1 评分函数的分解 …… 183
8.5.2 K2算法 …… 184
8.5.3 爬山法 …… 186
8.6 缺值数据结构学习 …… 188
8.6.1 SEM算法的基本思想 …… 188

8.6.2 SEM算法 …… 189
8.6.3 SEM的收敛性 …… 191
8.7 文献介绍 …… 192
第9章 隐结构模型学习 …… 194
9.1 隐变量与隐变量模型 …… 194
9.2 可分辨性及几个相关概念 …… 195
9.3 隐变量模型选择 …… 196
9.4 隐类模型 …… 197
9.4.1 正则性 …… 198
9.4.2 隐变量模型选择与大样本理论 …… 199
9.4.3 隐类模型学习算法 …… 200
9.4.4 仿真试验 …… 201
9.5 多层隐类模型 …… 204
9.5.1 走根运算与模型等价 …… 205
9.5.2 无根 HLC 模型 …… 206
9.5.3 正则性 …… 207
9.5.4 正则模型空间的有限性 …… 209
9.6 多层隐类模型学习算法 …… 211
9.6.1 势学习算法 …… 211
9.6.2 模型学习算法 …… 212
9.6.3 复杂度分析 …… 216
9.6.4 仿真试验 …… 217
9.7 文献介绍 …… 220

第四部分　贝叶斯网应用

第10章 隐结构模型与中医辨证 …… 225
10.1 中医辨证的客观化、定量化 …… 225
10.1.1 相关工作回顾 …… 225
10.1.2 隐结构法 …… 227
10.2 肾虚数据收集 …… 229
10.3 数据分析原理 …… 232
10.4 肾虚数据分析 …… 234
10.5 结果模型定性内容的质量 …… 236

10.6 结果模型定量内容的质量…… 239
10.7 结果模型与辨证论治…… 250
10.8 模型辨证的质量…… 252
10.9 讨论…… 260

参考文献…… 263
英汉词汇对照…… 278
索引…… 287

第一部分

贝叶斯网基础

第 1 章　概率论基础

本章简要介绍概率论中与贝叶斯网密切相关的一些基本概念. 与一般的概率论教材不同，我们将侧重于概念的直观含意. 深入理解诸如条件独立、贝叶斯定理等概念和结果，对以后理解贝叶斯网至关重要. 另外，本章还将简述概率方法在人工智能研究中的崛起过程，并介绍在第三部分中会被用到的一些信息论知识.

1.1　随机事件与随机变量

世界上许多事情都具有不确定性. 例如掷硬币，其结果可能正面朝上，也可能反面朝上，在抛掷之前无法预知. 又如赌马，理论上每匹马都有跑第一的可能，事先无法预料哪匹马一定会赢. 再如火星上是否曾有生命存在？答案有两种可能，是或不是，但根据目前掌握的证据判断，无法给出绝对的答复. 概率论是研究处理这类现象的数学理论. 本节介绍概率论中几个最基本的概念，包括样本空间、事件、概率测度、随机变量以及概率函数.

1. 样本空间和事件

在概率论中，**随机试验**指的是事先不能完全预知其结果的试验. 随机试验的所有可能结果组成该试验的**样本空间**，通常记为 Ω. 样本空间可以是离散的，也可以是连续的. 如无特殊说明，本书所论及的样本空间都将是离散的.

样本空间中的点，即随机试验的可能结果，称为**样本点**，或**原子事件**，记为 ω. 样本空间的子集称为**事件**，通常用大写字母表示：A，B，…. 如果随机试验的结果包含在一个事件之中，则称该事件发生了. 样本空间 Ω 本身也是一个事件，而且是一定会发生的**必然事件**. 空集也是一个事件，是**不可能事件**，通常记为 $\varnothing$. 事件之间可以进行交（$\cap$）、并（$\cup$）、差（$\setminus$）等各种集合运算. 若两事件 A 和 B 交空，即 $A \cap B = \varnothing$，则称它们为**互斥事件**，又称**不相容事件**. 两互斥事件不能同时发生. 若 A 和 B 互斥，且 $A \cup B = \Omega$，则称它们为**互补事件**.

例 1.1　考虑掷硬币试验，其结果有正反面两种可能，因此样本空间为 $\Omega = \{h, t\}$，其中 h 表示正面，t 表示反面，h 和 t 为两个互补的原子事件. □

例 1.2　考虑掷骰子试验，有 6 种可能的结果，样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$. 子集合 $\{1, 3, 5\}$ 表示的是“掷出结果为奇数”这一事件，其互补事件为 $\{2, 4, 6\}$，即“掷出结果为偶数”. □

例 1.3　考虑从香港科技大学的所有研究生中随机抽样的试验，其结果可能

是任一研究生，样本空间为 $\Omega = \{x \mid x$ 是科大研究生$\}$. {研究生张三} 为一原子事件①. {所有科大男研究生} 或 {所有年龄超过 25 岁的研究生} 是两个非原子事件. □

2. 概率测度

概率测度给样本空间中的每一事件 A 赋予一个数值 $P(A) \in [0,1]$，以度量该事件发生的可能性. 在数学上，它是一个从样本空间 Ω 的幂集 2^{Ω} 到区间 $[0, 1]$ 的映射 $P: 2^{\Omega} \to [0, 1]$，且满足以下 3 个 **Kolmogorov 公理**（Kolmogorov, 1933, 1950）：

(1) $P(\Omega) = 1$;

(2) $P(A) \geqslant 0$, $\forall A \in 2^{\Omega}$;

(3) $P(A \cup B) = P(A) + P(B), \forall A, B \in 2^{\Omega}, A \cap B = \varnothing$.

$P(A)$ 称为事件 A 的**概率**. 上面这 3 个公理分别被称为概率测度的**规范性**、**非负性**和**有限可加性**. 规范性规定必然事件的概率为 1，非负性规定概率不能为负，有限可加性规定互斥事件的并集的概率等于它们各自概率的和. 从这 3 个公理出发，可以推出概率测度的诸多基本性质和定理.

例 1.4 例 1.1 中的样本空间的所有子集是：Ω, $\varnothing$, $\{h\}$, $\{t\}$. 设所掷为均匀硬币，则有 $P(\Omega) = 1$, $P(\varnothing) = 0$, $P(h) = P(t) = 0.5$. 这里 $P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

例 1.5 在例 1.1 中，若所掷为不均匀硬币，且已知在 1000 次重复试验中正面出现 700 次，反面出现 300 次，可因此设 $P(\Omega) = 1$, $P(\varnothing) = 0$, $P(h) = 0.7$, $P(t) = 0.3$. 这里 $P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

例 1.6 在例 1.3 中，对 Ω 的任一子集 A，定义 $P(A) = \frac{|A|}{|\Omega|}$，其中 $|A|$ 是 A 中元素的个数，称为 A 的势. 那么，$P(\cdot)$ 满足概率定义的 3 个公理，是一个概率测度. □

3. 随机变量

随机变量是定义在样本空间 Ω 上的函数，通常用大写字母表示，如 X, Y, Z. 随机变量的取值随试验的结果而定，通常用小写字母表示，如 x, y, z. 随机变量 X 的所有可能取值的集合称为它的**值域**，也称**状态空间**，记为 Ω_X. 随机变量可以是离散的，也可以是连续的. **离散随机变量**的状态空间是离散的，包含有限个或无穷可数个状态. **连续随机变量**的状态空间是连续的，包含无穷不可数个

① 假设只有一个叫张三的研究生.

状态. 本书主要考虑离散随机变量.

例 1.7　在例 1.3 中，设 X 为“随机抽出的一个学生的性别”，则 X 是定义在“科大研究生”这个样本空间上的随机变量：$\Omega_X = \{m, f\}$，其中 m 表示男，f 表示女. □

例 1.8　同时掷两个质地均匀的硬币，其样本空间为 $\Omega = \{(h, h), (h, t), (t, h), (t, t)\}$. 对任一 $\omega \in \Omega$, $P(\{\omega\}) = 1/4$. 设 X 为“正面朝上的硬币个数”，那么 X 是定义在 Ω 上的一个随机变量：$X((h, h)) = 2$, $X((h, t)) = 1$, $X((t, h)) = 1$, $X((t, t)) = 0$, 故 $\Omega_X = \{0, 1, 2\}$. □

4. 概率函数

设 X 为一随机变量，x 是它的一个取值. 在样本空间中，所有使 X 取值为 x 的原子事件组成一个事件，记之为 $\Omega_{X=x} = \{\omega \in \Omega \mid X(w) = x\}$，有时也简记为“$X = x$”. 注意，$\Omega_{X=x}$ 与 Ω_X 的含意完全不同，后者是随机变量 X 的状态空间，包括 X 的所有可能取值.

事件“$X = x$”的概率 $P(X = x) = P(\Omega_{X=x})$ 依赖于 X 的取值 x. 让 x 在 Ω_X 上变动，$P(X = x)$ 就成为 Ω_X 的一个取值于 $[0,1]$ 的函数，称之为随机变量 X 的**概率质量函数**，记为 $P(X)$. 根据概率测度的定义，有

$$P(X = x) \geqslant 0, \forall x \in \Omega_X; \sum_{x \in \Omega_X} P(X = x) = 1.$$

为了记号上的方便，上面两式有时简记为

$$P(X) \geqslant 0; \quad \sum_X P(X) = 1.$$

例 1.9　在例 1.7 中，设科大共有 500 名研究生，其中 400 名男生，100 名女生. 这里 $\Omega_X = \{m, f\}$, $\Omega_{X=f} = \{$科大所有女研究生$\}$. $X = f$ 的概率为

$$P(X = f) = P(\Omega_{X=f}) = P(\{\text{科大所有女研究生}\}) = 100/500 = 0.2.$$

X 的概率函数为

X	m	f
$P(X)$	0.8	0.2

□

例 1.10　在例 1.8 中，变量 X 的值域是 $\Omega_X = \{0, 1, 2\}$, $\Omega_{X=0} = \{(t, t)\}$, $\Omega_{X=1} = \{(t, h), (h, t)\}$, $\Omega_{X=2} = \{(h, h)\}$. X 的概率函数为

X	0	1	2
$P(X)$	0.25	0.5	0.25

□

离散随机变量有概率质量函数. 与之对应，连续随机变量有概率密度函数. 由于本书主要关心离散随机变量，所以对概率密度函数不做详细介绍. 以后，我们有时会用“概率分布”一词来泛指概率质量函数、概率密度函数，或与它们等价的其它概念.

1.2 概率的解释

概率的解释主要有 5 种：古典解释、频率解释、主观解释、特性解释（Popper，1957）以及逻辑解释（Carnap，1950）. 本节主要介绍前 3 种解释，重点放在概率的主观解释上.

1.2.1 古典解释

概率的古典解释起源于 16 世纪数学家们对掷骰子等赌博活动的研究. 对一粒质地均匀的正方体骰子，投掷后其任何一面朝上的可能性相等，因此掷出每面的概率都应为 1/6. 一般地讲，如果事件 A 包含的样本数为 m，而样本空间的总样本数为 n，则事件 A 的概率应为

$$P(A)=\frac{\text{事件 } A \text{ 包含的样本数}}{\text{样本空间的总样本数}}=\frac{m}{n}. \tag{1.1}$$

用这种方法定义的概率称为**古典概率**. 古典概率的一个前提条件就是等可能性. 在实际应用中，这个前提一般很难满足，因此古典概率的应用范围很有限.

例 1.11 考虑如下掷 3 粒骰子的赌局：若结果之和为 {3，4，5，6，7，14，15,16，17，18} 中任一数字，则赌客胜，否则庄家胜. 3 粒骰子之和有从 3 ～ 18 共 16 种可能结果，其中赌客胜的有 10 种，所以表面看来似乎这是一个对赌客有利的赌局. 但是，由于结果不具有等可能性，因此不能简单认为赌客获胜的概率是 10/16. 实际上，这是一个对庄家有利的赌局，投掷 3 粒骰子有 216 种等可能结果：{(1，1，1)，(1，1，2)，…，(6，1，1)，…，(6，6，6)}，其中使赌客获胜的结果只有 69 种，而使庄家获胜的结果有 147 种. □

1.2.2 频率解释

给定一个质地不均匀的骰子，掷出 6 的概率为多大？古典解释无法处理这种情况，因为这时等可能性前提不成立. 为近似计算这一概率，人们通常进行多次重复试验，记下其中掷出 6 的次数，除之以总的试验次数，将结果作为掷出 6 的概率. 一般地讲，对于一个可在同样条件下重复进行的试验，如果事件 A 在所有 N 次试验中共发生了 M 次，则它的概率可以用其发生的频率来近似：$P(A)\approx M/N$. 这个近似的理论支持是**大数定律**：当 N 趋于无穷大时，频率几乎处处趋

于概率. 即当 N 较大时，频率经常稳定地出现在概率附近，而当 N 越大时，越是更经常地稳定于概率，而且幅度也越小. 这就是概率的**频率解释**.

按照频率解释，概率只有当试验可以在同等条件下无限次重复时才有意义. 然而，实际中人们往往需要研究一些不可重复的事件发生的概率，例如总统竞选或体育比赛的结果. 频率解释对这些一次性事件无法处理. 在早期的人工智能研究中，概率的频率解释曾占据主导地位，这一度为概率论的应用造成了概念上的困难.

1.2.3　主观解释

主观解释又称贝叶斯解释，它认为概率即合理**信度**，反映的是个体的知识状态和主观信念. 在这种意义下的概率称为**主观概率**.

1. *主观概率的评估*

相对于频率解释，主观解释的长处是它允许对一次性事件也进行概率评估. 例如：巴西队赢得下届世界杯足球赛冠军的概率是多大？频率解释认为此问题无意义，因为“下届世界杯决赛巴西队夺冠”不是一个可以重复的事件. 但是，主观解释仍可以根据各种先验知识给出一个主观概率评估. 主观概率的评估有许多方法，其中之一是下面的概率轮方法.

例 1.12（概率轮与概率评估）　设想有一质地均匀的概率轮（图 1.1），其上仅包含黑白两个连续区域，转动后指针停在任一位置的概率相等，因而其停在黑区的概率应等于黑区角度所占的百分比. 概率轮提供了一个进行主观概率评估的客观参照. 当评估“巴西队赢得下届世界杯足球赛冠军的概率”时，首先问如下问题：巴西队夺冠的可能性大，还是指针停在黑区的可能性大？如果认为巴西队夺冠更有可能，那么就设想一个更大的黑区，反之设想一个较小的黑区，再次问同样的问题. 如此反复，直到认为巴西队夺冠和指针停在黑区具有相同的可能性，然后测量黑区角度的大小，除以 360，即得到巴西队夺冠的概率. □

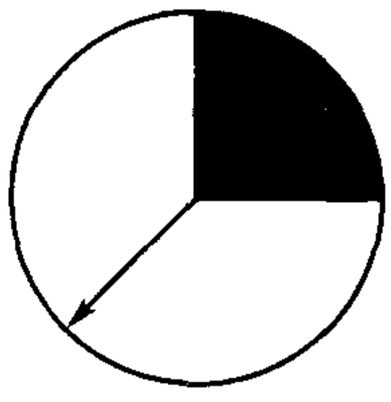


图 1.1　用于主观概率评估的概率轮

作为一种操作性强的手段，用概率轮进行主观概率评估在管理科学、心理学以及运筹学中被广泛使用，详见有关文献（Spetzler and von Holstein，1975）. 不难看出，这里存在一个评估精度的问题：概率值为 0.201 或 0.202，有什么根本的差别吗？两者差别如此之小，以至于人们往往很难断定自己的真实信度究竟是哪一个. 所幸的是，在贝叶斯网应用中，这往往不是一个大问题. 原因有三：首先，概率值的微小差别对决策的影响一般不大；其次，实际中往往会同时考虑多个事件的

概率，由于概率必须满足 Kolmogorov 公理，因此不同事件的概率之间存在一定的关系，而这些关系限制了主观概率的任意性；第三，在数据分析中，当数据量足够大时，主观概率的影响不大，这一点将在本小节最后详细论述.

2. 主观概率与 Kolmogorov 公理

为什么主观概率必须满足 Kolmogorov 公理呢？围绕这个问题，人们提出了几种理论和论证方法（Shafer and Pearl，1990）. 其中之一与赌博有关，它的基本思想是：如果一个人的主观概率不满足 Kolmogorov 公理，那么就可以构造一个赌局，使他认为合理而接受，但又将必输无疑，这样的赌局称为 Dutch book①. 对于**理性个体**，不应该有 Dutch book 存在，因此理性个体的主观概率必须满足 Kolmogorov 公理.

如果一个人对某事件 S 出现的主观概率为 P，$0 \leqslant P \leqslant 1$，则他最多愿意付 P 元去购买一个关于 S 的单位筹码，即如果 S 为真则可拿它兑换 1 元赔金的筹码. 同时，他也愿意以 P 元或更高的价格卖掉一个单位筹码. 下面的两个例子分别说明，如果一个人的主观概率不满足 Kolmogorov 第(1)和第(3)公理，那么就有针对他的 Dutch book 存在.

例 1.13 考虑 $\Omega = \{S, \neg S\}$. 假设一赌客的主观概率为 $P(S) = 0.51$，$P(\neg S) = 0.51$，从而总和 $P(\Omega) = P(S) + P(\neg S)$ 大于 1，违反了 Kolmogorov 第（1）公理. 因为 $P(S) = 0.51$，所以他愿意以 0.51 元买入一个赌 S 为真的单位筹码 C_S；同时因为 $P(\neg S) = 0.51$，所以他也愿意以 0.51 元买入一个赌 $\neg S$ 为真的单位筹码 $C_{\neg S}$. 构造如下的 Dutch book：让赌客以 1.02 元的价格同时买进 C_S 和 $C_{\neg S}$ 两个筹码. 对这个赌客来讲，此 Dutch book 是公平的. 但是一旦他接受，无论结果是 S 还是 $\neg S$，他都只能拿回 1 元，从而损失 0.02 元. □

例 1.14 考虑一个由 3 匹马 H_1，H_2，H_3 参加的跑马比赛. 假设有 8 种彩票 T_0，T_1，T_2，T_3，T_{12}，T_{13}，T_{23}，T_{123}，它们的赔金如下：

T_1：100 元 如果 H_1 赢	T_{12}：100 元 如果 H_1 或 H_2 赢
T_2：100 元 如果 H_2 赢	T_{13}：100 元 如果 H_1 或 H_3 赢
T_3：100 元 如果 H_3 赢	T_{23}：100 元 如果 H_2 或 H_3 赢
T_0：100 元 如果没有任何一匹马赢	T_{123}：100 元 如果任何一匹马赢

分别以 $P(H_1)$，$P(H_2)$，$P(H_1 \cup H_2)$ 记某人对 H_1 赢、H_2 赢、H_1 或 H_2 赢

① 在英式赌马比赛中，一个“book”是指为某人所接受的一套下注组合.

的主观概率估计，那么对他来说，彩票 T_1、T_2 和 T_{12} 的公平价格分别为 $P(H_1)\times 100$，$P(H_2)\times 100$ 和 $P(H_1\cup H_2)\times 100$.

假设一赌客的概率评估如下：$P(H_1)=0.3$，$P(H_2)=0.4$，$P(H_1\cup H_2)=0.5$. 这里 $P(H_1)+P(H_2)\neq P(H_1\cup H_2)$，违反了 Kolmogorov 第(3)公理. 一个针对此赌客的 Dutch book 如下：假设他手中开始时有彩票 T_{12}，从他手中以 50 元买走 T_{12}，再分别以 30 元和 40 元的价格将 T_1 和 T_2 卖给他. 对他来说，这个交易是公平的，但是他却蒙受了损失. 交易前后赌客手中彩票的价值变化如下：

	交易前	交易后
如果 H_1 赢	100 元	100＋50－30－40＝80（元）
如果 H_2 赢	100 元	100＋50－30－40＝80（元）
如果 H_3 赢	0 元	50－30－40＝－20（元）

无论哪匹马赢，该赌客都损失 20 元. □

3. 主观概率与贝叶斯网

贝叶斯网早期主要应用于专家系统. 在专家系统应用中，贝叶斯网的结构和参数是通过咨询专家而获得的，因此需要用类似于概率轮的方法进行概率评估，主观概率占有重要地位.

随着时间的推移，贝叶斯网越来越多地被用于分析数据，也就是要基于数据建立贝叶斯网模型. 这有两种情形：一是已知网络结构，对网络参数进行估计，称为参数学习；二是不知道网络结构，要通过分析数据，同时获得网络结构和网络参数，称为结构学习. 参数学习有两种方法——最大似然估计和贝叶斯估计. 最大似然估计完全基于数据，不需要先验概率. 贝叶斯估计则假定在考虑数据以前，网络参数服从某个先验分布. 这是先验的主观概率，它的影响随着数据量的增大而减小. 结构学习情形假设在考虑数据以前，不同结构的可能性相等，这也是先验的主观概率，它的影响也随着数据量的增大而减小. 所以，当有足够多的数据时，主观概率对数据分析的影响不大.

尽管概率的主观解释在贝叶斯网的实际应用中并不扮演非常重要的角色，但是在概念上，它对贝叶斯网却是至关重要的. 贝叶斯网所依赖的一个核心概念是条件独立，而概率的主观解释为直观理解条件概率和条件独立提供了一个自然的角度. 这一点的论证将在第 1.3.3 和 1.3.4 节中看到.

1.2.4　特性解释与逻辑解释

在特性解释中，均匀硬币“正面朝上”的概率为 1/2 是这个硬币的固有物理属性，与其是否投掷或投掷次数无关. 特性解释没有为概率提供可操作的运算方

法，因此很难应用于实际之中.

逻辑解释则认为概率是对知识状态的总结，是由从证据到假设的逻辑关系所决定的. 一旦相关的知识得到确定，则事件的可能性就已经被客观地确定下来，并应该能够通过逻辑分析来得到. 古典解释可以看做是逻辑解释的一个特例，它从等可能性的前提条件出发来计算概率. 同特性解释一样，逻辑解释的缺点在于它没能为概率提供一个可操作的运算方法.

1.3 多元概率分布

随机现象往往涉及多个随机因素，因而需要用多个随机变量来描述. 本节介绍多元概率的一些基本概念.

1.3.1 联合概率分布

我们知道，对单个随机变量 X，可以用概率函数 $P(X)$ 来描述它的各个状态的概率. 而对于多个随机变量 $X_1, \cdots, X_n$，则可以用**联合概率分布** $P(X_1, \cdots, X_n)$，简称**联合分布**来描述各变量所有可能的状态组合的概率. 它是一个定义在所有变量状态空间的笛卡儿乘积之上的函数：

$$P(X_1, \cdots, X_n): \bigotimes_{i=1}^{n} \Omega_{X_i} \to [0, 1],$$

其中所有函数值之和为 1，即

$$\sum_{X_1, \cdots, X_n} P(X_1, \cdots, X_n) = 1.$$

联合分布经常被表示为一张表，其中包含了 $\prod_{i=1}^{n} |\Omega_{X_i}|$ 个状态组合及其概率值. 如果所有变量都只取两个状态，则联合分布表共有 2^n 个项，刻画了变量之间的各种关系.

例 1.15 考虑香港市场上所有出租房屋. 从中随机抽取一间，考查其月租(R)和类型(T)这两个随机变量. 月租分为 4 等：{low(低于 2000 元)，medium (2000 ~ 6000 元)，upper medium (6000 ~ 12000 元)，high(高于 12000 元)}. 类型有 3 种：{public(公屋)，private(私家屋)，others(其它)}. 联合分布 $P(R, T)$ 如下：

R \ T	public	private	others
low	0.17	0.01	0.02
medium	0.44	0.03	0.01
upper medium	0.09	0.07	0.01
high	0	0.14	0.01

从表中可知，随机抽到中价公屋的可能性最大，为 44%. □

1.3.2 边缘概率分布

在例 1.15 中，由于有了联合分布 $P(R,T)$，所以可以回答这样的问题：随机抽取一间出租房屋为公屋的概率 $P(T=\text{public})$ 是多少？根据概率的有限可加性

$$\begin{aligned}P(T=\text{public}) &= P(T=\text{public}, R=\text{low})\\ &\quad + P(T=\text{public}, R=\text{medium})\\ &\quad + P(T=\text{public}, R=\text{upper medium})\\ &\quad + P(T=\text{public}, R=\text{high})\\ &= 0.7.\end{aligned}$$

同样地，可以计算 $P(T=\text{private})$ 和 $P(T=\text{others})$：

$$\begin{aligned}P(T=\text{private}) &= P(T=\text{private}, R=\text{low})\\ &\quad + P(T=\text{private}, R=\text{medium})\\ &\quad + P(T=\text{private}, R=\text{upper medium})\\ &\quad + P(T=\text{private}, R=\text{high})\\ &= 0.25,\\ P(T=\text{others}) &= P(T=\text{others}, R=\text{low})\\ &\quad + P(T=\text{others}, R=\text{medium})\\ &\quad + P(T=\text{others}, R=\text{upper medium})\\ &\quad + P(T=\text{others}, R=\text{high})\\ &= 0.05.\end{aligned}$$

为了简化记号，上面三式可分别缩写为

$$P(T=\text{public}) = \sum_R P(T=\text{public}, R),$$

$$P(T=\text{private}) = \sum_R P(T=\text{private}, R),$$

$$P(T=\text{others}) = \sum_R P(T=\text{others}, R).$$

这三个式子还可以进一步合并为下面一式：

$$P(T) = \sum_R P(T, R).$$

相对于联合分布 $P(R, T)$，$P(T)$ 称为**边缘分布**①. 下表同时给出了联合分布 $P(R, T)$ 和边缘分布 $P(T)$，$P(R)$：

① 这一术语来源于保险统计行业. 保险统计师通常把观察到的频率数据相加，并把结果写在保险统计报表的边缘，所以叫“边缘概率”.

R \ T	public	private	others	$P(R)$
low	0.17	0.01	0.02	0.20
medium	0.44	0.03	0.01	0.48
upper medium	0.09	0.07	0.01	0.17
high	0	0.14	0.01	0.15
$P(T)$	0.70	0.25	0.05	

记 $\mathbf{X}=\{X_1, \cdots, X_n\}$，$\mathbf{Y}$ 是 $\mathbf{X}$ 的真子集，即 $\mathbf{Y}\subset\mathbf{X}$，$\mathbf{Z}=\mathbf{X}\backslash\mathbf{Y}$. 则相对于 $P(\mathbf{X})$，$\mathbf{Y}$ 的边缘分布 $P(\mathbf{Y})$ 定义为

$$P(\mathbf{Y})=\sum_{\mathbf{Z}} P(X_1, \cdots, X_n). \tag{1.2}$$

从联合分布 $P(\mathbf{X})$ 到边缘分布 $P(\mathbf{Y})$ 的过程称为**边缘化**.

例 1.16 设有 3 个装有黑白两色球的口袋，第 1 个口袋黑白球各半，第 2 个口袋黑白球比例为 4∶1，第 3 个则全是黑球. 设随机变量 X, Y, Z 分别代表从这 3 个口袋随机抽出的球的颜色，其状态空间为 $\Omega_X=\Omega_Y=\Omega_Z=\{w, b\}$，其中 w 表示白，b 表示黑. 联合概率分布 $P(X,Y,Z)$ 如下：

X	Y	Z	$P(X,Y,Z)$
w	w	w	0
w	w	b	0.1
w	b	w	0
w	b	b	0.4
b	w	w	0
b	w	b	0.1
b	b	w	0
b	b	b	0.4

表中给出了 X, Y, Z 的所有 8 个可能的状态组合及其概率. 从表中可知抽球结果为 (w, b, b) 和 (b, b, b) 的概率一样大，都是 0.4；结果为 (w, w, b) 和 (b, w, b) 的概率也一样，都是 0.1；而其余所有满足 $Z=w$ 的状态组合的概率都为零，因为第 3 个袋子里没有白球. 这里边缘分布 $P(X)$ 为 (0.5, 0.5)，$P(Y)$ 为 (0.2,0.8)，$P(Z)$ 为 (0,1). □

1.3.3　条件概率分布

条件概率与条件分布是用来刻画事件之间及变量之间关系的基本工具.

1. 条件概率

设 A, B 为两随机事件且 $P(B)>0$，事件 A 在给定事件 B 发生时的**条件概率**定义为

$$P(A \mid B)=\frac{P(A \cap B)}{P(B)}. \tag{1.3}$$

直观上，$P(A \mid B)$ 是在已知 B 发生时，对 A 发生的信度；而 $P(A)$ 则是在不知道 B 是否发生时，对 A 发生的信度①. 由式 (1.3) 可得

$$P(A \cap B)=P(B) P(A \mid B). \tag{1.4}$$

这称为**概率的乘法定律**，其含意非常直观：对 A 和 B 同时发生的信度，等于对 B 发生的信度乘以已知 B 发生时对 A 发生的信度. 当然乘法定律也可以写为

$$P(A \cap B)=P(A) P(B \mid A). \tag{1.5}$$

例 1.17　在例 1.2 的掷骰子试验中，掷出 6 的概率为 1/6. 假定投掷后被告知“掷出的结果是偶数”，问此时对结果为 6 的信度是多少？设掷出 6 为事件 A，掷出结果为偶数为事件 B，则 $P(A)=1/6$，$P(B)=1/2$，$P(A \cap B)=1/6$. 所问的问题即是要计算如下条件概率：

$$P(A \mid B)=\frac{P(A \cap B)}{P(B)}=\frac{1/6}{1/2}=\frac{1}{3}.$$

□

2. 条件分布

设 X 和 Y 是两随机变量，x 和 y 分别是它们的一个取值. 考虑事件 $X=x$ 在给定 $Y=y$ 时的条件概率为

$$P(X=x \mid Y=y)=\frac{P(X=x, Y=y)}{P(Y=y)}. \tag{1.6}$$

在式(1.6)中，固定 y，让 x 在 Ω_X 上变动，则得到一个在 Ω_X上的函数. 这个函数称为**在给定 $Y=y$ 时变量 X 的条件概率分布**，记为 $P(X \mid Y=y)$. 用 $P(X \mid Y)$ 记 $\{P(X \mid Y=y) \mid y \in \Omega_Y\}$，即在 Y 取不同值时 X 的条件概率分布的集合. $P(X \mid Y)$ 称为**给定 Y 时变量 X 的条件概率分布**. 在式(1.6)中，让 x 和 y 在 Ω_X

① 这里用到的是概率的主观解释.

和 Ω_Y 上变动，则得到一组等式. 与 1.3.2 节类似，把这些等式缩写为

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}. \tag{1.7}$$

式 (1.7) 可视为是 $P(X \mid Y)$ 的直接定义.

更一般地，设 $\mathbf{X} = \{X_1, \cdots, X_n\}$ 和 $\mathbf{Y} = \{Y_1, \cdots, Y_m\}$ 为两个变量集合，$P(\mathbf{X},\mathbf{Y})$ 为 $\mathbf{X} \cup \mathbf{Y}$ 的联合概率分布，$P(\mathbf{Y})$ 为 $\mathbf{Y}$ 的边缘概率分布. 则给定 $\mathbf{Y}$ 时 $\mathbf{X}$ 的条件概率分布定义为

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{X},\mathbf{Y})}{P(\mathbf{Y})}. \tag{1.8}$$

例 1.18 在例 1.15 中问：随机抽取一间私家屋，其租金为 low 的概率多大? 这即是问给定 $T=$private 时 $R=$low 的条件概率 $P(R=\text{low} \mid T=\text{private})$. 按定义，有

$$P(R = \text{low} \mid T = \text{private}) = \frac{P(R = \text{low}, T = \text{private})}{P(T = \text{private})} = \frac{0.01}{0.25} = 0.04.$$

给定 T 时，变量 R 的条件分布 $P(R \mid T)$ 如下：

T \ R	low	medium	upper medium	high
public	$\frac{0.17}{0.7}$	$\frac{0.44}{0.7}$	$\frac{0.09}{0.7}$	$\frac{0}{0.7}$
private	$\frac{0.01}{0.25}$	$\frac{0.03}{0.25}$	$\frac{0.07}{0.25}$	$\frac{0.14}{0.25}$
others	$\frac{0.02}{0.05}$	$\frac{0.01}{0.05}$	$\frac{0.01}{0.05}$	$\frac{0.01}{0.05}$

表中第 1 行显示的是在给定 $T=$public 时，R 的条件概率分布，第 2 行是在给定 $T=$ private 时，R 的条件概率分布，等等. 这里每行的数字之和为 1，即 $\sum_R P(R \mid T) = 1$. 这与例 1.15 中所列的联合分布 $P(R, T)$ 不同，那里表中所有数字之和为 1，即 $\sum_{R,T} P(R, T) = 1$. □

下面的例子涉及本节前面所介绍的 3 个主要概念，即联合分布、边缘分布和条件分布. 其目的是为读者建立这样的直观印象：几个随机变量的联合分布对应的是一张表，其中一些变量的边缘分布以及条件分布对应的也是表.

例 1.19 设有一袋积木，每块积木有 3 个属性：颜色、材料和形状. 设积木的颜色只能是红(r)或蓝(b)两种，材料只能是金属(m)或木头(w)，形状只能是正方体(6)或正四面体(4). 设 C, M, S 为 3 个随机变量，分别代表从袋中随机取出一块积木的颜色、材料和形状，则 $\Omega_C = \{r, b\}$, $\Omega_M = \{m, w\}$, $\Omega_S =$

{6, 4}. 设联合概率分布 $P(C, M, S)$ 为

C	M	S	$P(C,M,S)$
r	m	6	0.10
r	m	4	0.10
r	w	6	0.25
r	w	4	0.05
b	m	6	0.15
b	m	4	0.10
b	w	6	0.20
b	w	4	0.05

那么，变量 C 和 M 的边缘分布 $P(C, M)$ 为

C	M	$P(C,M)$
r	m	0.20
r	w	0.30
b	m	0.25
b	w	0.25

条件分布 $P(C \mid M)$ 为

M \ C	r	b
m	$\frac{4}{9}$	$\frac{5}{9}$
w	$\frac{6}{11}$	$\frac{5}{11}$

□

3. 链规则

对两变量 X、Y 的联合分布 $P(X,Y)$，按照条件分布的定义，可得

$$P(X,Y) = P(X)P(Y \mid X). \tag{1.9}$$

将其推广到 n 个变量的联合分布 $P(X_1, X_2, \cdots, X_n)$，有

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2 \mid X_1) \cdots P(X_n \mid X_1, \cdots, X_{n-1}). \tag{1.10}$$

式(1.10)将一个联合分布分解为一系列条件分布的乘积，它称为**链规则**.

1.3.4 边缘独立与条件独立

1. 事件独立

设 A, B 为同一随机试验的两个不同事件. 我们称**事件 A 与 B 相互独立**，如果下式成立：

$$P(A \cap B) = P(A)P(B), \tag{1.11}$$

当 $P(B) > 0$ 时，由式 (1.11) 可得 $P(A) = P(A \mid B)$. $P(A \mid B)$ 是已知事件 B 发生时对 A 发生的信度，而 $P(A)$ 是在未知事件 B 是否发生时对 A 发生的信度. 所以 A 与 B 相互独立的直观含义是：对于事件 B 是否发生的了解不影响对事件 A 发生的信度. 当 $P(A) > 0$ 时，由式 (1.11) 可得 $P(B) = P(B \mid A)$，所以 A 与 B 相互独立意味着：对于事件 A 是否发生的了解也不影响对事件 B 发生的信度.

考虑 3 个事件 A, B 和 C，假定 $P(C) > 0$. 我们称**事件 A 与 B 在给定 C 时相互条件独立**，如果下式成立：

$$P(A \cap B \mid C) = P(A \mid C)P(B \mid C), \tag{1.12}$$

当 $P(B \cap C) > 0$ 时，由式 (1.12) 可得 $P(A \mid C) = P(A \mid B \cap C)$. $P(A \mid C)$ 是已知事件 C 发生时对事件 A 发生的信度，而 $P(A \mid B \cap C)$ 是已知 事件 B 和 C 都已发生时对事件 A 发生的信度. 所以，事件 A 与 B 在给定 C 时相互条件独立的直观意义是：在已知事件 C 发生的前提下，对事件 B 是否发生的了解不会改变对事件 A 发生的信度；同样，对事件 A 是否发生的了解也不影响对事件 B 发生的信度.

2. 变量独立

两个随机变量 X 和 Y 称为**相互（边缘）独立**，记为 $X \perp Y$，如果下式成立：

$$P(X, Y) = P(X)P(Y), \tag{1.13}$$

考虑变量 Y 的某个取值 y，如果 $P(Y = y) > 0$，则由式 (1.13) 可得

$$P(X) = P(X \mid Y = y).$$

$P(X \mid Y = y)$ 是已知 $Y = y$ 时，变量 X 的概率(信度)分布，而 $P(X)$ 是未知 Y 的取值时 X 的概率(信度)分布. 所以，变量 X 与 Y 相互独立意味着：对变量 Y 的取值的了解不会改变变量 X 的概率(信度)分布；同样，对变量 X 的取值的了解也不会改变变量 Y 的概率(信度)分布.

更一般地，我们称随机变量 X_1, X_2, …, X_n 相互（边缘）独立，如果

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2)\cdots P(X_n).$$

例 1.20 在例 1.16 中，从 3 个袋子中抽球，所得球的颜色的联合概率分布

$P(X,Y,Z)$ 如下：

X	Y	Z	$P(X,Y,Z)$
w	w	w	0
w	w	b	0.1
w	b	w	0
w	b	b	0.4
b	w	w	0
b	w	b	0.1
b	b	w	0
b	b	b	0.4

而边缘分布 $P(X)$，$P(Y)$ 和 $P(Z)$ 则分别如下：

X	w	b
$P(X)$	0.5	0.5

Y	w	b
$P(Y)$	0.2	0.8

Z	w	b
$P(Z)$	0	1

容易验证 $P(X,Y,Z)=P(X)P(Y)P(Z)$，即 X,Y,Z 相互边缘独立. □

考虑 3 个随机变量 X,Y 和 Z，设 $P(Z=z)>0,\forall z\in\Omega_Z$. 我们说 X 和 Y **在给定 Z 时相互条件独立**，记为 $X\perp Y\mid Z$. 如果下式成立：

$$P(X,Y\mid Z)=P(X\mid Z)P(Y\mid Z). \tag{1.14}$$

设 y 和 z 分别是 Y 和 Z 的任意取值，且 $P(Y=y,\ Z=z)>0$，由式（1.14）可得

$$P(X\mid Y=y,\ Z=z)=P(X\mid Z=z).$$

$P(X\mid Z=z)$ 是在已知 $Z=z$ 时，X 的概率（信度）分布，而 $P(X\mid Y=y,Z=z)$ 是在已知 $Y=y$ 以及 $Z=z$ 时，X 的概率（信度）分布. 因此，$X\perp Y\mid Z$ 的直观含意是：在已知 Z 的前提下，对 Y 的取值的了解不影响 X 的概率（信度）分布. 注意，这并不意味着在未知 Z 的取值时，X 和 Y 相互独立. $Y=y$ 有可能含有关于 X 的信息，只是所有这样的信息也都包含于 $Z=z$ 中，所以当已知 $Z=z$ 时，进一步了解到 $Y=y$ 并不增加关于 X 的信息. 当然，$X\perp Y\mid Z$ 也意味着，在已知 Z 的取值时，对 X 的取值的了解不影响 Y 的概率（信度）分布.

例 1.21　设有一装有两种硬币的口袋，其中一些是均匀硬币，掷出正面朝上的概率为 0.5；另一些为非均匀硬币，掷出正面朝上的概率为 0.8. 现从袋中随机取出一个硬币，抛掷若干次. 令 X_i 表示第 i 次抛掷硬币的结果，Y 表示该硬币

是否均匀. 这里，X_i 与 X_j $(i \neq j)$ 之间不是相互（边缘）独立的，因为如果掷了 10 次硬币，其中 9 次都是正面朝上，那么有理由相信这枚硬币是不均匀的，从而增大了下一次掷出正面朝上的信度. 所以 X_i 的值给了我们关于这枚硬币的一些信息，它有助于我们继续判断 X_j 的值.

另一方面，如果已经知道了 Y 的值，例如该硬币是不均匀的，那么不管前面的结果如何，以后每次掷硬币的结果为正面的概率都是 0.8，我们将不能从前面的试验得到什么信息. 所以给定 Y 的值后，X_i 与 X_j 之间就是相互条件独立的. 本例中变量间的依赖关系可以用图 1.2 来表示：变量 Y 切断了变量 X_i 与变量 X_j 之间的“信息通道”. □

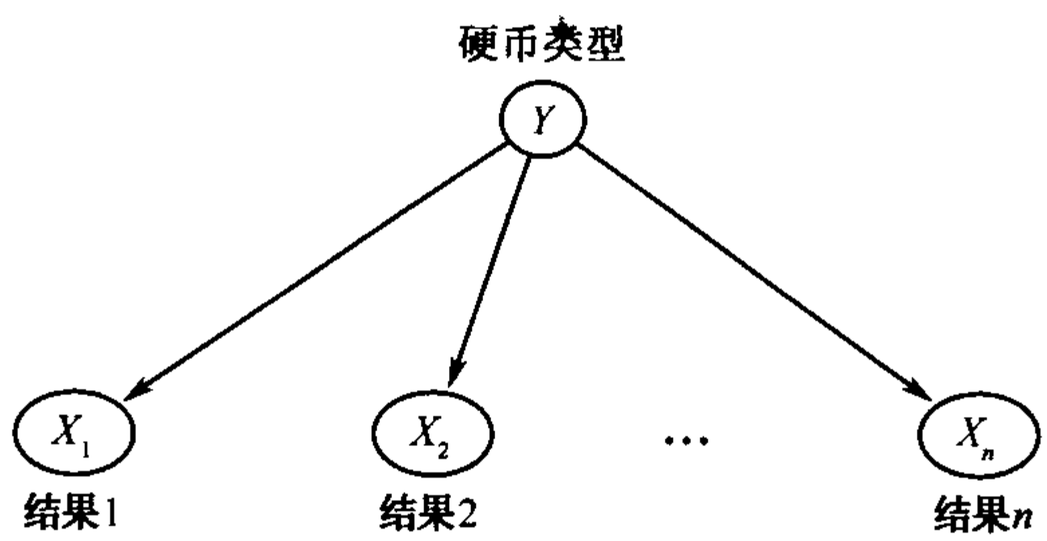


图 1.2 条件独立示意：给定硬币类型，各投掷结果相互独立

命题 1.1 考虑 3 个随机变量 X, Y 和 Z，设 $P(Z) > 0$，下列条件相互等价：

(1) $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$；

(2) $P(X \mid Y, Z) = P(X \mid Z)$，当 $P(Y, Z) > 0$；

(3) $P(X, Y \mid Z) = f(X, Z)g(Y, Z)$，$f$ 和 g 均为函数；

(4) $P(X \mid Y, Z) = f(X, Z)$，f 为一函数，当 $P(Y, Z) > 0$；

(5) $P(X, Y, Z) = P(X \mid Z)P(Y \mid Z)P(Z)$；

(6) $P(X, Y, Z) = \dfrac{P(X, Z)P(Y, Z)}{P(Z)}$；

(7) $P(X, Y, Z) = f(X, Z)g(Y, Z)$，$f$ 和 g 均为函数.

1.3.5 贝叶斯定理

先验概率和后验概率这两个概念是相对于某组证据而言的. 设 H 和 E 为两个随机变量，$H = h$ 为某一假设，$E = e$ 为一组证据. 在考虑证据 $E = e$ 之前，对事件 $H = h$ 的概率估计 $P(H = h)$ 称为**先验概率**. 而在考虑证据之后，对 $H = h$ 的概率估计 $P(H = h \mid E = e)$ 称为**后验概率**. **贝叶斯定理**描述了先验概率和后验概率之间的关系：

$$P(H = h \mid E = e) = \frac{P(H = h)\ P(E = e \mid H = h)}{P(E = e)}, \tag{1.15}$$

这又称为**贝叶斯规则**，或**贝叶斯公式**.

例 1.22　假设有两个装满糖果的口袋 a 和 b. 袋 a 中有 10 块水果糖和 30 块巧克力糖. 袋 b 中水果糖和巧克力糖各有 20 块. 随机选择一个口袋，并从中随机取出一块糖果，发现是巧克力糖. 问这块糖是从袋 a 中抽出的可能性为多大?

用 C 表示取出的糖块，H 表示它来自的口袋，T 表示它的类型. 这里所关心的假设是 $H=a$，而证据是 $T=$“巧克力糖”. 在考虑证据之前，C 来自袋 a 和袋 b 的概率相同，都是 0.5，所以先验概率是 $P(H=a)=0.5$. $P(T=\text{“巧克力糖”}\mid H=a)$ 是在已知 C 来自袋 a 的情况下它为巧克力的概率，即

$$P(T=\text{“巧克力糖”}\mid H=a)=30/40=0.75.$$

由于总共有 80 块糖，其中的 50 块是巧克力，所以

$$P(T=\text{“巧克力糖”})=50/80=0.625.$$

利用贝叶斯公式，有

$$\begin{aligned}P(H=a\mid T=\text{“巧克力糖”})&=\frac{P(H=a)P(T=\text{“巧克力糖”}\mid H=a)}{P(T=\text{“巧克力糖”})}\\&=\frac{0.5\times 0.75}{0.625}=0.6.\end{aligned}$$

这就是 $H=a$ 的后验概率. □

在贝叶斯定理中，$P(E=e\mid H=h)$ 称为 $H=h$ 的**似然度**，有时记为 $L(H=h\mid E=e)$. 贝叶斯定理之所以有用是因为似然度往往容易获得，而后验概率则不然. 在例 1.22 中，似然度 $P(T=\text{“巧克力糖”}\mid H=a)$ 由袋中不同糖果的组成所决定，是容易得到的；而后验概率 $P(H=a\mid T=\text{“巧克力糖”})$ 则要通过一番推理计算才能得到. 下面再来看一个例子.

例 1.23　设有一病人眼睛呈黄色，医生要判断他患乙肝的可能性. 这里的假设是病人患有乙肝 $D=t$，证据是病人眼睛呈黄色 $C=y$. 要直接判断 $P(D=t\mid C=y)$ 并不容易. 但根据对乙肝的了解，医生知道乙肝病人眼黄的概率是 80%，即似然度 $P(C=y\mid D=t)=0.8$；而在医生的所有病人中，0.5% 的人患有乙肝，10%的人眼黄，即先验概率 $P(D=t)=0.005$，而 $P(C=y)=0.1$. 利用贝叶斯公式，有

$$\begin{aligned}P(D=t\mid C=y)&=\frac{P(D=t)P(C=y\mid D=t)}{P(C=y)}\\&=\frac{0.005\times 0.8}{0.1}=0.04.\end{aligned}$$

即病人患乙肝的概率是 0.04，这是先验概率 0.005 的 8 倍. □

相对于证据 $E=e$，可以谈论一个事件的先验和后验概率，同时也可以谈论一个变量的先验和后验概率分布. 设 X 是一个所关心的变量，则有

$$P(X\mid E=e)=\frac{P(X)P(E=e\mid X)}{P(E=e)}.\tag{1.16}$$

这是贝叶斯定理的变量形式. 式中 $P(X)$ 是 X 的**先验分布**，$P(X \mid E=e)$ 是 X 的**后验分布**，$P(E=e \mid X)$ 称为 X 的**似然函数**，有时记为 $L(X \mid E=e)$. $P(E=e)$ 是一个**规一化常数**，它保证式子右边的函数是一个概率分布，其值由下式定义：

$$P(E=e)=\sum_{X} P(X)\ P(E=e \mid X).$$

式（1.16）中，$P(E=e)$ 不依赖 X，所以有时把它写为如下形式：

$$P(X \mid E=e) \propto P(X) L(X \mid E=e), \tag{1.17}$$

即后验分布正比于先验分布和似然函数的乘积.

1.4 概率论与人工智能

贝叶斯网是为了处理人工智能研究中的不确定性问题而发展起来的. 人工智能兴起于 20 世纪 50 年代中期，它的目标是研究人类智能的机理，提供智能行为的计算模型，进而构造能具有智能行为的系统. 专家系统是人工智能的一个子领域，它的目标是将某一复杂领域的专家的知识和经验引入计算机系统，使得更多人能够借助计算机系统受惠于专家的经验.

不确定性是人工智能系统所面临的一个重要问题. 它来源于多个方面. 首先，智能系统对外部环境的观测往往是不完备的或是有误差的. 其次，推理涉及对复杂世界一定程度的抽象和简化，因此推理的前提往往因为例外情况而得不到完全满足. 在实际应用中，推理前提的例外很多，不能穷举，不确定性提供了一个总结各种例外情况的机制. 第三，多数复杂领域并没有一套完备的理论，从而推理必须在不确定中进行. 最后，有些客观规律本身就是统计的、随机的、非确定性的.

在六七十年代人工智能的早期研究中，基于规则的专家系统占据主导地位. 这类系统又称为产生式系统，它的知识库由多个“IF-THEN”语句构成，使用一阶逻辑和谓词演算进行符号推理. 由于产生式系统在早期人工智能研究中取得了较大成就，所以人们最先想到的是对它进行简单扩充以处理不确定性. 有些学者致力于发明各种新的逻辑，用非单调推理①的方式来处理推理前提的例外，这些逻辑包括：默认逻辑（Reiter，1987）、非单调模态逻辑（McDermott and Doyle，1987）、自认知逻辑（Moore，1987）、限定逻辑（McCarthy，1980）等.

---

① 在高等数学中函数的单调性指的是如下特性：如果 $x \geqslant y$，那么 $f(x) \geqslant f(y)$. 相似地，一阶谓词演算也具有如下单调性：如果 A 是一个规则集合，则随着新规则的加入，它所能推出的结论集合也单调增大. 因此一阶逻辑通常被称为是单调逻辑，相应的推理叫单调推理. 可是，日常生活中人们的常识推理却是非单调的：我们通常直接“跳入”某一结论，而后随着新事实的发现再对原来以为为真的结论进行修正. 这种推理叫非单调推理（Ginsberg，1987）.

另一些学者则试图对产生式系统进行简单修补，通过为推理规则附加一个数字来量化不确定性. 对这个附加数字的语义解释有多种，包括确定因子（Shortliffe，1976）、概率数（Duda et al.，1976），以及模糊集合论中的隶属度（Zadeh，1965）等. 针对不同的解释，又有各种不同的演算规则. 这些方法在实际应用中都有各种各样的困难，主要原因是计算复杂度太高和推理结果的正确性不能保证. 有关详细讨论请读者参见有关文献（Heckerman，1985；Pearl，1988；Shafer and Pearl，1990）.

概率论是研究随机性的数学. 用概率论处理不确定性的主要优点是能够保证推理结果的正确性. 但是，在 20 世纪 80 年代以前，人们普遍认为概率论不适用于人工智能的研究，这有几个原因. 首先，**频率学派**占支配地位，他们认为概率只有在试验可以在同等条件下无限次重复时才有意义，这就排除了对一次性事件谈论概率的可能性. 其次，人们还没有想到利用问题的结构对联合概率分布进行分解，所以使用概率的计算复杂度太高，难以实现. 第三，不少学者认为数值运算是计算机的专长，而符号推理才是人类智能的独有特点，因此符号逻辑，而非概率数值运算，才是研究人工智能的恰当框架. 另外，一些学者则往往倚重一些启发式方法和编程技巧来实现某种程度的智能行为.

自 80 年代以来，概率论重新引起了人们的兴趣，逐渐成为处理人工智能中不确定性的主流方法. 这有三方面的原因. 首先，到 70 年代末，越来越多的人开始认同概率的贝叶斯解释，将概率当做人的主观信念程度来看待，这使得在频率数据缺乏的情况下也可以使用概率（Shafer and Pearl，1990）. 其次，**决策论**（Von Neumann and Morgenstern，1947）与人工智能的结合对概率论的复兴起到了巨大的推动作用. 决策论的一个核心法则是**最大期望效用**原则，基于决策论的智能系统被称为规范型系统. 与之相对，早期的人工智能系统侧重于描述和总结专家解决问题的逻辑推理过程或所使用的启发式规则，称为**描述型系统**. 两者的主要区别在于：①描述型系统是对专家建模，规范型系统是对问题领域建模；②描述型系统所使用的不确定性运算规则不能保证推理结果的正确性，规范型系统则使用概率运算规则和决策论，可以得到正确的结果；③描述型系统强调替代专家决策，规范型系统强调支持专家决策(Jensen,1996). 第三，在 80 年代初，有学者发现，利用问题的结构可以把联合概率分布进行分解，从而大大降低计算复杂度(Pearl,1982；Kim and Pearl，1983；Pearl，1986)，这对概率方法在人工智能研究中的崛起起到了关键性的作用.

1.5　信息论基础

信息论是建筑于概率论之上的研究信息传输和信息处理的数学理论. 它不仅

是信息技术的基础，还在诸如统计力学、机器学习等其它领域中起着重要作用. 本书的第三部分将会利用信息论的一些结果来对贝叶斯网学习算法进行分析. 同时，信息论也有助于加深读者对于“条件独立”这一概念的理解. 本节扼要介绍信息论的一些基本概念和结果.

1.5.1 Jensen 不等式

一个函数 f 在实数轴的某个区间 I 上被称为**凹函数**，如果 $\forall x_1, x_2 \in I$，有

$$f(\lambda x_1 + (1-\lambda)x_2) \geqslant \lambda f(x_1) + (1-\lambda) f(x_2), \forall \lambda \in [0,1]. \quad (1.18)$$

这一关系如图 1.3(a)所示. 若式（1.18）中的等号只在 $x_1 = x_2$ 时才成立，则称 f 在区间 I 上**严格凹**. 如果将式（1.18）中的不等号改变方向，就得到**凸函数**的定义，如图 1.3(b)所示①. 如果 f 是凹函数，则 $-f$ 是凸函数. 图 1.4 给出了信息论中几个常见的凹函数.

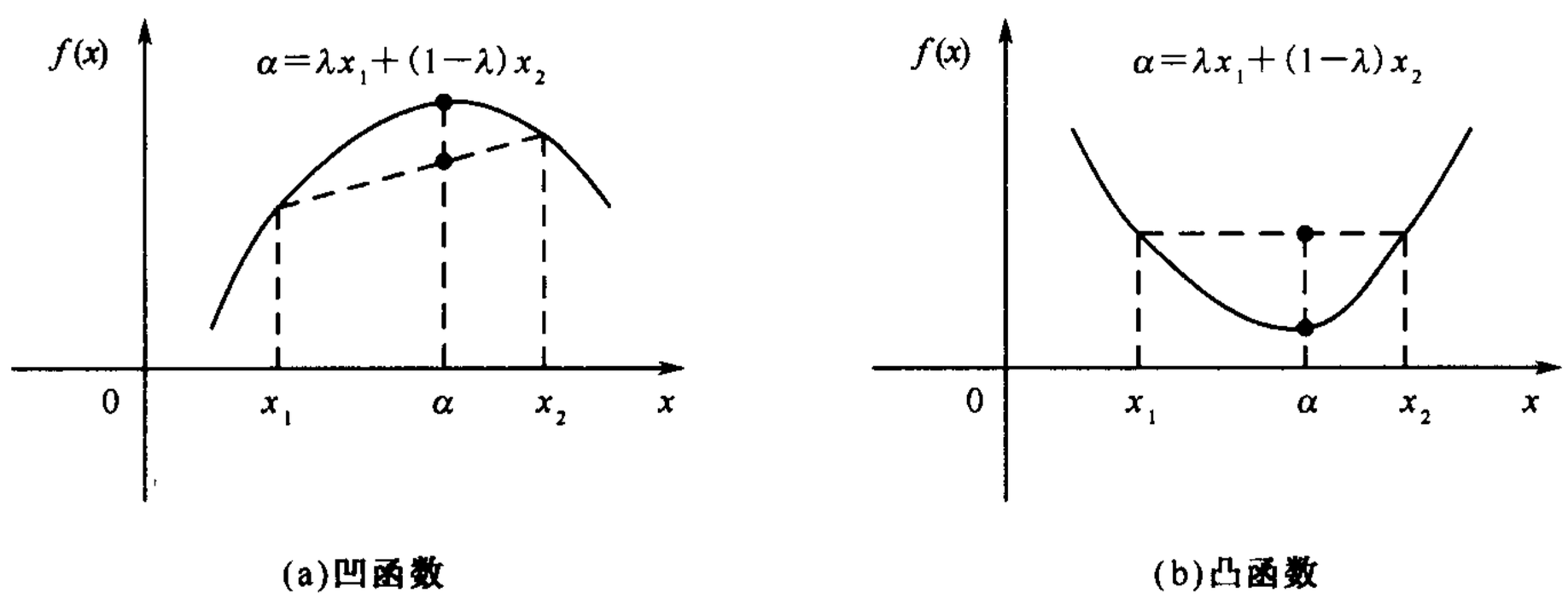


图 1.3 凹函数和凸函数

定理 1.1（Jensen 不等式） 设 f 为区间 I 上的凹函数，$p_i \in [0, 1], i = 1, 2, \cdots, n$，且 $\sum_{i=1}^{n} p_i = 1$. 则对任何 $x_i \in I$，有

$$f\left(\sum_{i=1}^{n} p_i x_i\right) \geqslant \sum_{i=1}^{n} p_i f(x_i). \quad (1.19)$$

若 f 严格凹，则式（1.19）的等号只在下列条件满足时才成立：若 $p_i \cdot p_j \neq 0$，则必有 $x_i = x_j$.

证明 用归纳法证明. 当 $n = 1$ 时，式（1.19）显然成立. 假设式（1.19）

① 注意，凸函数和凹函数的数学曲线与汉字中“凸”和“凹”的直观形状刚好相反. 英文中凸函数是“convex”，对应开口向上的下凸曲线“∪”；凹函数是“concave”，对应开口向下的上凸曲线“∩”. 这一点经常容易引起混淆.

在 $n=k$ 时成立，证明它在 $n=k+1$ 时也成立，即

$$
\begin{aligned}
f\left(\sum_{i=1}^{k+1} p_i x_i\right) &= f\left(\sum_{i=1}^{k} p_i x_i + p_{k+1} x_{k+1}\right) \\
&= f\left[(1-p_{k+1})\frac{1}{1-p_{k+1}}\sum_{i=1}^{k} p_i x_i + p_{k+1} x_{k+1}\right] (\text{假设 } p_{k+1} \neq 1) \\
&\geqslant (1-p_{k+1})\, f\left(\frac{1}{1-p_{k+1}}\sum_{i=1}^{k} p_i x_i\right) + p_{k+1} f(x_{k+1}) \quad (\text{根据凹函数定义}) \\
&= (1-p_{k+1}) f\left(\sum_{i=1}^{k} \frac{p_i}{1-p_{k+1}} x_i\right) + p_{k+1} f(x_{k+1}) \\
&\geqslant (1-p_{k+1}) \sum_{i=1}^{k} \frac{p_i}{1-p_{k+1}} f(x_i) + p_{k+1} f(x_{k+1}) \quad (\text{根据归纳假设}) \\
&= \sum_{i=1}^{k} p_i f(x_i) + p_{k+1} f(x_{k+1}). \\
&= \sum_{i=1}^{k+1} p_i f(x_i).
\end{aligned}
$$

式（1.19）得证．我们把 f 严格为凹的部分的证明作为练习留给读者．　□

Jensen 不等式是凹函数的基本性质，在信息论中经常用到．也有关于凸函

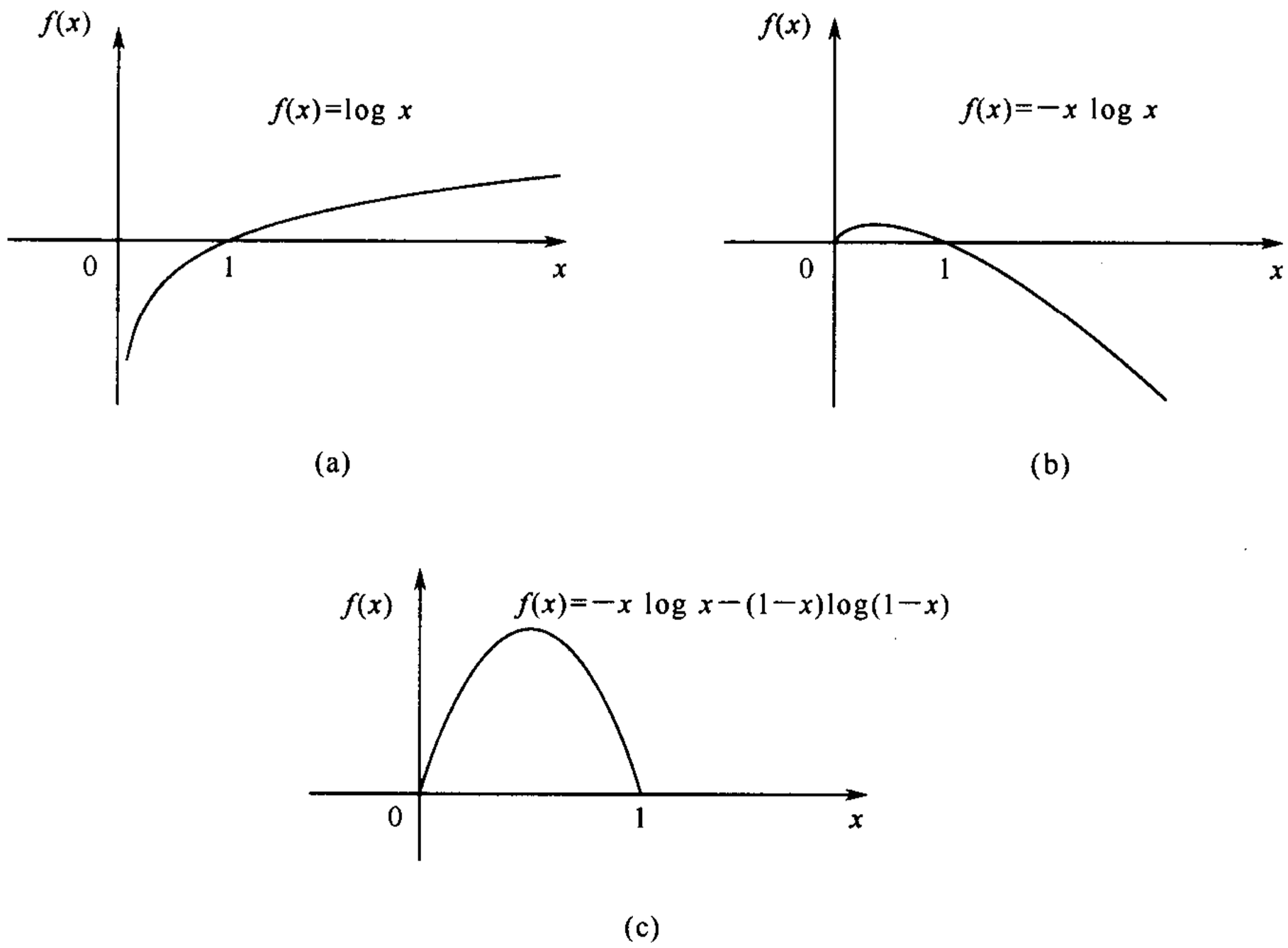


图 1.4　信息论中常见的凹函数

数的 Jensen 不等式，它可以通过将式（1.19）中的不等号换个方向而得到.

例 1.24 图 1.4(a)所示，对数函数 $f(x)=\log x$ 在区间 $(0,+\infty)$ 是凹函数. 根据 Jensen 不等式，有

$$\log\left(\sum_{i=1}^{n} p_i x_i\right) \geqslant \sum_{i=1}^{n} p_i \log x_i. \tag{1.20}$$

式（1.20）中，$\forall i, x_i>0$，$p_i \geqslant 0$，且 $\sum_{i=1}^{n} p_i = 1$. □

1.5.2 熵

一个离散随机变量 X 的**熵** $H(X)$ 的定义为

$$H(X)=\sum_{X} P(X)\log\frac{1}{P(X)}=-\sum_{X} P(X)\log P(X), \tag{1.21}$$

其中约定 $0\log\frac{1}{0}=0$. 式（1.21）的对数若以 2 为底，则熵的单位是**比特**；若以 e 为底，则其单位是**奈特**. 若无特殊说明，本书以后章节均采用比特为单位.

熵是对随机变量的不确定性的度量. 随机变量 X 的熵越大，说明它的不确定性也越大. 下面举两个例子来说明这一点.

例 1.25 考虑一个取值为 0 或 1 的随机变量 X，记 $p=P(X=1)$. 根据熵的定义，有

$$H(X)=-p\log p-(1-p)\log(1-p).$$

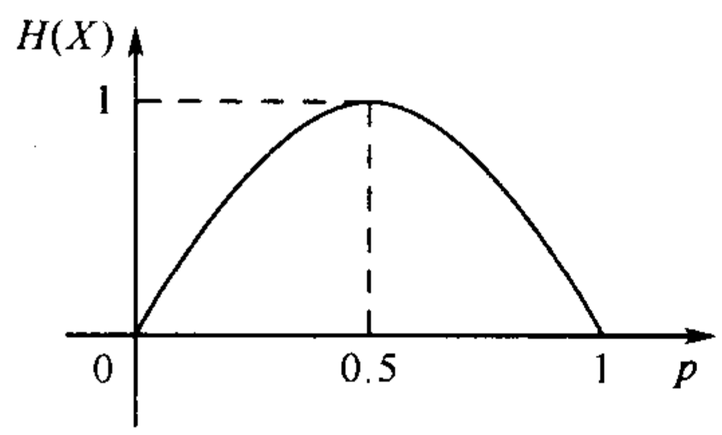


图 1.5 二值随机变量的熵，这里 $p=P(X=1)$

这个函数如图 1.5 所示. 当 $p=0$ 或 $p=1$ 时，我们肯定地知道 X 的取值，不确定性最小，$H(X)=0$. 当 $p=0.5$ 时，相当于投掷一个质地均匀的硬币，对 X 的取值的不确定性达到最大，此时 $H(X)=1$. □

例 1.26 记 X，Y 和 Z 分别为掷硬币、掷骰子以及从 54 张扑克牌中随意抽取一张的结果. 显然 X 的不确定性最小，Y 的不确定性居中，而 Z 的不确定性最大. 与之相应，这 3 个随机变量的熵之间也恰恰存在这样的关系，即

$$H(X)<H(Y)<H(Z);$$

$$H(X)=\frac{1}{2}\log 2+\frac{1}{2}\log 2=\log 2=1;$$

$$H(Y)=\underbrace{\frac{1}{6}\log 6+\cdots+\frac{1}{6}\log 6}_{6\text{个}}=\log 6;$$

$$H(Z)=\underbrace{\frac{1}{54}\log 54+\cdots+\frac{1}{54}\log 54}_{54\text{ 个}}=\log 54.$$

□

用 $|X|$ 来记变量 X 的取值个数，又称为变量的**势**.

命题 1.2（熵的基本性质）

(1) $H(X)\geqslant 0$;

(2) $H(X)\leqslant \log|X|$，等号成立当且仅当对 X 的所有取值 x 有 $P(X=x)=\frac{1}{|X|}$ ①.

证明　(1) 显然成立，因为对于 X 的任意取值 x，总有

$$-P(X=x)\log P(X=x)\geqslant 0.$$

对于 (2)，根据式 (1.20)，有

$$\begin{aligned}H(X)&=\sum_X P(X)\log\frac{1}{P(X)}\\&\leqslant \log\sum_X P(X)\frac{1}{P(X)}=\log|X|.\end{aligned}$$

命题得证. □

1.5.3　联合熵、条件熵和互信息

联合熵是借助联合概率分布对熵的自然推广. 两个离散随机变量 X 和 Y 的联合熵的定义为

$$H(X,Y)=\sum_{X,Y}P(X,Y)\log\frac{1}{P(X,Y)}=-\sum_{X,Y}P(X,Y)\log P(X,Y). \tag{1.22}$$

条件熵是利用条件概率分布对熵的一个延伸. 随机变量 X 的熵是用它的概率分布 $P(X)$ 来定义的. 如果知道另一个随机变量 Y 的取值为 y，那么 X 的后验分布即为 $P(X\mid Y=y)$. 利用此条件分布可以定义给定 $Y=y$ 时 X 的条件熵为

$$H(X\mid Y=y)=\sum_X P(X\mid Y=y)\log\frac{1}{P(X\mid Y=y)}. \tag{1.23}$$

熵 $H(X)$ 度量的是随机变量 X 的不确定性，条件熵 $H(X\mid Y=y)$ 度量的则是已知 $Y=y$ 后，X 的不确定性.

在式(1.23)中，当 y 变化时，$H(X\mid Y=y)$ 也会发生改变. 由于知道 Y 的概率分布，因此可以计算观测 Y 后 X 的熵的期望值，即

---

① 这个性质经常被称为**最大熵原理**.

$$
\begin{aligned}
H(X \mid Y) &= \sum_{y \in \Omega_Y} P(Y=y)\ H(X \mid Y=y) \\
&= \sum_{y \in \Omega_Y} P(Y=y) \sum_X P(X \mid Y=y) \log \frac{1}{P(X \mid Y=y)} \\
&= \sum_Y \sum_X P(Y) P(X \mid Y) \log \frac{1}{P(X \mid Y)} \\
&= \sum_{X,Y} P(X, Y) \log \frac{1}{P(X \mid Y)}.
\end{aligned} \tag{1.24}
$$

$H(X \mid Y)$ 称为给定 Y 时 X 的条件熵.

注意 $H(X \mid Y)$ 与 $H(X \mid Y=y)$ 有所不同. 后者是在已知 Y 取某一特定值 y 时 X 的条件熵，或者说是在已知 $Y=y$ 后，X 剩余的不确定性. 而 $H(X \mid Y)$ 则是在未知 Y 的取值时，对观测到 Y 的取值后 X 剩余的不确定性的一个期望. 尤其值得注意的是，$H(X \mid Y=y)$ 可能比 $H(X)$ 大，即知道 Y 的具体取值 $Y=y$ 可能增大对 X 的不确定性；但读者在 1.5.5 节将看到，$H(X \mid Y)$ 永远不大于 $H(X)$，即平均来说，知道 Y 将不会增加 X 的不确定性.

例 1.27 设联合分布 $P(X,Y)$ 以及边缘分布 $P(X)$ 和 $P(Y)$ 如下：

	x_1	x_2	$P(Y)$
y_1	0	$\frac{3}{4}$	$\frac{3}{4}$
y_2	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
$P(X)$	$\frac{1}{8}$	$\frac{7}{8}$	

从而可以得出

$$H(X) = -\frac{1}{8}\log\frac{1}{8} - \frac{7}{8}\log\frac{7}{8} = 0.544;$$

$$H(X \mid Y=y_1) = -0\log 0 - 1\log 1 = 0;$$

$$H(X \mid Y=y_2) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1;$$

$$H(X \mid Y) = \frac{3}{4}H(X \mid Y=y_1) + \frac{1}{4}H(X \mid Y=y_2) = 0.25.$$

可以看到，观测到 $Y=y_1$ 使 X 的熵减小，观测到 $Y=y_2$ 使 X 的熵增大. 但平均来说，对 Y 的观测使 X 的熵减小. □

在观测到 Y 以前，X 的不确定性是 $H(X)$. 通过观测 Y，我们期望 X 的不确定性会变为 $H(X \mid Y)$. 因此 $H(X)$ 与 $H(X \mid Y)$ 之差

$$I(X;Y) = H(X) - H(X \mid Y) \tag{1.25}$$

就是对 Y 包含多少关于 X 的信息的一个度量，称之为 Y 关于 X 的信息. 很快我们将看到 $I(X;Y) = I(Y;X)$，因此它又称为 X 和 Y 之间的**互信息**. 以下定理给出联合熵、相对熵和互信息之间的重要关系.

定理 1.2　对任意两个离散随机变量 X 和 Y，有

$$I(X;Y) = \sum_{X,Y} P(X,Y)\log\frac{P(X,Y)}{P(X)P(Y)}, \tag{1.26}$$

$$I(X;Y) = I(Y;X), \tag{1.27}$$

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y), \tag{1.28}$$

$$I(X;Y) + H(X,Y) = H(X) + H(Y). \tag{1.29}$$

其中式 (1.28) 称为**熵的链规则**.

证明　(1) 对式 (1.26)，有

$$\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= \sum_{X} P(X)\log\frac{1}{P(X)} - \sum_{X,Y} P(X,Y)\log\frac{1}{P(X \mid Y)} \\
&= \sum_{X,Y} P(X,Y)\log\frac{1}{P(X)} - \sum_{X,Y} P(X,Y)\log\frac{1}{P(X \mid Y)} \\
&= \sum_{X,Y} P(X,Y)\log\frac{P(X \mid Y)}{P(X)} \\
&= \sum_{X,Y} P(X,Y)\log\frac{P(X,Y)}{P(X)P(Y)}.
\end{aligned}$$

(2) 对式 (1.27)，这是第(1)步结果的显然推论.

(3) 对式 (1.28)，有

$$\begin{aligned}
H(X,Y) &= -\sum_{X,Y} P(X,Y)\log P(X,Y) \\
&= -\sum_{X,Y} P(X,Y)\log P(X) - \sum_{X,Y} P(X,Y)\log P(Y \mid X) \\
&= -\sum_{X} P(X)\log P(X) - \sum_{X,Y} P(X,Y)\log P(Y \mid X) \\
&= H(X) + H(Y \mid X).
\end{aligned}$$

同理可证 $H(X,Y) = H(Y) + H(X \mid Y)$.

(4) 对式 (1.29)，有

$$\begin{aligned}
I(X;Y) + H(X,Y) &= (H(X) - H(X \mid Y)) + (H(Y) + H(X \mid Y)) \\
&= H(X) + H(Y).
\end{aligned}$$

定理得证. □

联合熵、条件熵和互信息之间的关系可以用如图 1.6 所示的**韦恩图**来总结.

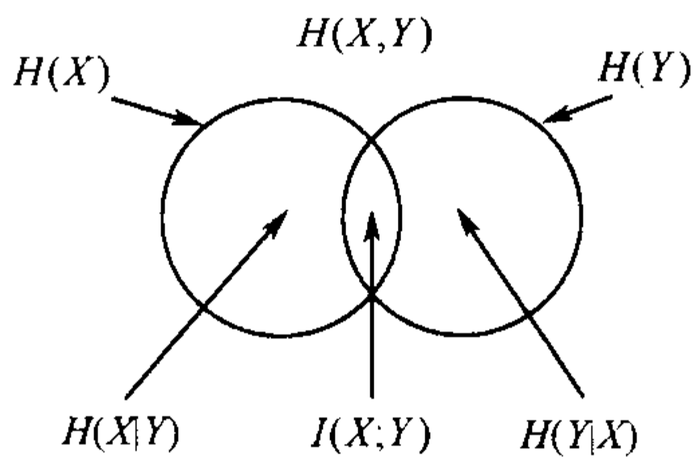


图 1.6 联合熵、条件熵以及互信息之间的关系

1.5.4 相对熵

对定义于随机变量 X 的状态空间 Ω_X 上的两个概率分布 $P(X)$ 和 $Q(X)$，可以用**相对熵**来度量它们之间的差异，即有

$$KL(P,Q)=\sum_X P(X)\log\frac{P(X)}{Q(X)}. \tag{1.30}$$

其中约定：$0\log\frac{0}{q}=0$；$p\log\frac{p}{0}=\infty,\forall p>0$. $KL(P,Q)$ 又被称为 $P(X)$ 和 $Q(X)$ 之间的 **Kullback-Leibler 距离**．但严格来讲它不是一个真正意义上的距离，因为 $KL(P,Q)\neq KL(Q,P)$.

定理 1.3（信息不等式） 设 $P(X)$ 和 $Q(X)$ 为定义在某个变量 X 的状态空间 Ω_X 上的两个概率分布，则有

$$KL(P,\ Q)\geqslant 0, \tag{1.31}$$

其中，当且仅当 P 与 Q 相同，即 $P(X=x)=Q(X=x)$，$\forall x\in\Omega_X$ 时等号成立.

证明

$$\begin{aligned}\sum_X P(X)\log\frac{P(X)}{Q(X)}&=-\sum_X P(X)\log\frac{Q(X)}{P(X)}\\&\geqslant-\log\sum_X P(X)\frac{Q(X)}{P(X)}\text{（根据式(1.20)）}\\&=-\log\sum_X Q(X)\\&=-\log 1=0.\end{aligned}$$

定理得证. □

以下推论在本书的第三部分将会被多次用到.

推论 1.1 对于满足 $\sum_X f(X)>0$ 的非负函数 $f(X)$，定义概率分布 $P^*(X)$ 为

$$P^*(X)=\frac{f(X)}{\sum_X f(X)}.$$

那么对于任意其它的概率分布 $P(X)$，则有

$$\sum_X f(X)\log P^*(X)\geqslant\sum_X f(X)\log P(X),$$

其中当且仅当 P^* 与 P 相同时等号成立.

证明　根据定理 1.3，有

$$KL(P^*,P)=\sum_X P^*(X)\log\frac{P^*(X)}{P(X)}\geqslant 0.$$

因此有

$$\sum_X P^*(X)\log P^*(X)\geqslant\sum_X P^*(X)\log P(X),$$

即

$$\sum_X\frac{f(X)}{\sum_X f(X)}\log P^*(X)\geqslant\sum_X\frac{f(X)}{\sum_X f(X)}\log P(X).$$

从而有

$$\sum_X f(X)\log P^*(X)\geqslant\sum_X f(X)\log P(X).$$

推论得证. □

1.5.5　互信息与变量独立

本节指出互信息与变量独立之间的两个关系. 首先有以下定理.

定理 1.4　对任意两个离散随机变量 X 和 Y，有

(1) $I(X;Y)\geqslant 0$;

(2) $H(X\mid Y)\leqslant H(X)$.

上面两式当且仅当 X 与 Y 相互独立时等号成立.

证明　由式 (1.26) 可得

$$I(X;Y)=KL(P(X,Y),P(X)P(Y)),\tag{1.32}$$

即 $I(X;Y)$ 是分布于 $P(X,Y)$ 和 $P(X)P(Y)$ 之间的相对熵. 根据信息不等式，$I(X;Y)\geqslant 0$，当且仅当 $P(X,Y)=P(X)P(Y)$ 时等号成立. 亦即 $I(X;Y)=0$ 当且仅当 X 与 Y 相互独立. 由于 $I(X;Y)=H(X)-H(X\mid Y)$，所以 $H(X\mid Y)\leqslant H(X)$，而且 $H(X\mid Y)=H(X)$ 当且仅当 X 与 Y 相互独立. 定理得证. □

定理 1.4 从信息论角度为“边缘独立”这一概念提供了一个直观解释，即两个随机变量相互独立当且仅当它们之间的互信息为零.

接下来考虑 3 个变量 X, Y 和 Z 之间的条件独立关系. 条件熵 $H(X \mid Z)$ 是给定 Z 时 X 剩余的不确定性. 如果进一步再给定 Y, X 剩余的不确定性变为 $H(X \mid Z, Y)$. 因此这两者之差即是给定 Z 时观测 Y 的取值会带来的关于 X 的信息量，即

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Z, Y), \tag{1.33}$$

称为给定 Z 时 Y 关于 X 的信息. 容易证明 $I(X;Y \mid Z) = I(Y;X \mid Z)$，于是 $I(X;Y \mid Z)$ 也称为给定 Z 时 X 和 Y 之间的**条件互信息**.

定理 1.5 对任意 3 个离散随机变量 X, Y 和 Z，有

(1) $I(X;Y \mid Z) \geqslant 0$;

(2) $H(X \mid Y,Z) \leqslant H(X \mid Z)$.

上面两式当且仅当 $X \perp Y \mid Z$ 时等号成立.

证明 从条件互信息的定义出发，有

$$\begin{aligned}
I(X;Y \mid Z) &= H(X \mid Z) - H(X \mid Y,Z) \\
&= \sum_{X,Z} P(X, Z) \log \frac{1}{P(X \mid Z)} - \sum_{X,Y,Z} P(X, Y, Z) \log \frac{1}{P(X \mid Y, Z)} \\
&= \sum_{X,Y,Z} P(X, Y, Z) \log \frac{1}{P(X \mid Z)} - \sum_{X,Y,Z} P(X, Y, Z) \log \frac{1}{P(X \mid Y, Z)} \\
&= \sum_{X,Y,Z} P(X, Y, Z) \log \frac{P(X \mid Y, Z)}{P(X \mid Z)} \\
&= \sum_{Z} P(Z) \sum_{X,Y} P(X, Y \mid Z) \log \frac{P(X, Y \mid Z)}{P(X \mid Z) P(Y \mid Z)} \\
&= \sum_{Z} P(Z) KL(P(X, Y \mid Z), P(X \mid Z) P(Y \mid Z)) \\
&\geqslant 0. \text{（根据定理 1.3）}
\end{aligned}$$

这样就同时证明了上面两式(1)和(2)，而且其中当且仅当 $P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$，即 $X \perp Y \mid Z$ 时等号成立. □

定理 1.5 的意义在于，它从信息论角度为随机变量之间的“条件独立”这一概念提供了一个直观解释，即给定 Z，两个随机变量 X 和 Y 相互条件独立，当且仅当它们的条件互信息为零. 或者说，Y 关于 X 的信息已全部包含在 Z 中，从而观测到 Z 后，再对 Y 进行的观测不会带来关于 X 的更多信息. 另一方面，如果 X 和 Y 在给定 Z 时相互不独立，则 $H(X \mid Z,Y) < H(X \mid Z)$，即在已知 Z 的基础上对 Y 的进一步观测将会带来关于 X 的新信息，从而降低 X 的不确定性.

第2章 贝叶斯网

本章介绍贝叶斯网的基本概念和定义. 我们首先分析直接用联合分布进行概率推理的困难，然后讨论如何利用变量间的条件独立关系将联合分布加以分解、降低模型表达和推理运算的复杂度，进而引出贝叶斯网的定义. 接着我们介绍如何手工构造贝叶斯网. 最后我们给出贝叶斯网的诸多应用实例，并介绍它对其它领域的影响.

2.1 不确定性推理与联合概率分布

不确定性推理是人工智能研究的重要课题之一. 从20世纪六七十年代以来，人们提出了多种方法，如概率方法（Nilsson，1965；Duda and Hart，1973）、非单调逻辑（Ginsberg，1987）、确定因子（Shortliffe and Buchanan，1975）、Dempster-Shafer证据理论（Shafter，1976）、模糊逻辑（Zadeh，1965）等. 在这些方法中，概率方法是最自然也是最早被尝试的方法之一，因为概率论本身是关于随机现象和不确定性的数学理论.

使用概率方法进行不确定性推理就是：①把问题用一组随机变量 $\mathbf{X}=\{X_1, X_2, \cdots, X_n\}$ 来刻画；②把关于问题的知识表示为一个联合概率分布 $P(\mathbf{X})$；③按照概率论原则进行推理计算. 下面来看一个例子.

例 2.1（Alarm 问题） Pearl 教授家住洛杉矶，那里地震和盗窃时有发生. 教授的家里装有警铃，地震和盗窃都可能触发警铃. 听到警铃后，两个邻居 Mary 和 John 可能会打电话给他. 一天，Pearl 教授接到 Mary 的电话，说听到他家警铃响，Pearl 教授想知道他家遭盗窃的概率是多大?

这个问题包含5个随机变量：盗窃（B）、地震（E）、警铃响（A）、接到 John 的电话（J）和接到 Mary 的电话（M）；所有变量的取值均是“y”或“n”. 这里各变量间的关系存在不确定性：盗窃和地震以一定概率随机发生；它们发生后，并不一定会触发警铃；而警铃响后，Mary 和 John 可能会因为某些原因，如在听摇滚乐或听力问题，而没有听到警铃；有时候，两人也会将其它声音误听为警铃声. 假设 Pearl 教授对这5个变量的联合概率分布 $P(B, E, A, J, M)$ 的评估如表2.1所示. 要计算的是接到 Mary 的电话（$M=y$）后，Pearl 教授对家里遭盗（$B=y$）的信度，即 $P(B=y \mid M=y)$.

表 2.1 Alarm 问题的联合概率分布 $P(B,E,A,J,M)$

B	E	A	M	J	概率	B	E	A	M	J	概率
y	y	y	y	y	1.2E-4	n	y	y	y	y	3.6E-3
y	y	y	y	n	5.1E-5	n	y	y	y	n	1.6E-3
y	y	y	n	y	1.3E-5	n	y	y	n	y	4.0E-4
y	y	y	n	n	5.7E-6	n	y	y	n	n	1.7E-4
y	y	n	y	y	5.0E-9	n	y	n	y	y	7.0E-6
y	y	n	y	n	4.9E-7	n	y	n	y	n	6.9E-4
y	y	n	n	y	9.5E-8	n	y	n	n	y	1.3E-4
y	y	n	n	n	9.4E-6	n	y	n	n	n	1.3E-2
y	n	y	y	y	5.8E-3	n	n	y	y	y	6.1E-4
y	n	y	y	n	2.5E-3	n	n	y	y	n	2.6E-4
y	n	y	n	y	6.5E-4	n	n	y	n	y	6.8E-5
y	n	y	n	n	2.8E-4	n	n	y	n	n	2.9E-5
y	n	n	y	y	2.9E-7	n	n	n	y	y	4.8E-4
y	n	n	y	n	2.9E-5	n	n	n	y	n	4.8E-2
y	n	n	n	y	5.6E-6	n	n	n	n	y	9.2E-3
y	n	n	n	n	5.5E-4	n	n	n	n	n	9.1E-1

从联合概率分布 $P(B, E, A, J, M)$ 出发，先计算边缘分布

$$P(B, M) = \sum_{E, A, J} P(B, E, A, J, M).$$

得到下表：

B	M	$P(B, M)$
y	y	0.000115
y	n	0.000075
n	y	0.00015
n	n	0.99966

再按照条件概率定义，得

$$\begin{aligned} P(B=y \mid M=y) &= \frac{P(B=y, M=y)}{P(M=y)} \\ &= \frac{P(B=y, M=y)}{P(B=y, M=y)+P(B=n, M=y)} \\ &= \frac{0.000115}{0.000115+0.000075} \approx 0.61. \end{aligned}$$

□

直接使用联合分布进行不确定性推理的困难很明显，即它的复杂度极高. 例2.1中有5个二值随机变量，整个联合分布包含 $2^5-1=31$ 个独立参数. 一般地，n 个二值变量的联合概率分布包含 2^n-1 个独立参数. 所以，联合分布的复杂度相对于变量的个数成指数增长. 当变量很多时，联合概率的获取、存储和运算都变得十分困难. 正是由于这个原因，在六七十年代，概率论被大多数学者认为不适合于解决人工智能中的不确定性问题.

2.2 条件独立与联合分布的分解

自20世纪80年代以来，概率方法重新在人工智能和专家系统的研究中获得重视，并逐渐占据主导地位. 这主要是因为人们发现，利用变量间的条件独立关系可以将联合分布分解成多个复杂度较低的概率分布，从而降低模型表达的复杂度，提高推理效率，使得人们可以应用概率方法来解决大型问题.

在例2.1中，运用链规则，可以把联合概率分布 $P(B,E,A,J,M)$ 表示为

$$\begin{aligned}&P(B,E,A,J,M)\\&=P(B)P(E\mid B)P(A\mid B,E)\,P(J\mid B,E,A)P(M\mid B,E,A,J).\end{aligned}\tag{2.1}$$

这一步并没有降低模型的复杂度，因为等式右端的5个概率分布仍然包含同联合分布相同个数的独立参数：$1+2+4+8+16=31$. 但是，它使得我们可以根据问题的背景知识做一些合理的独立假设以降低复杂度. 例如，地震（E）应与盗窃（B）无关，于是假设 E 与 B 相互独立，即 $P(E\mid B)=P(E)$，这样就把 $P(E\mid B)$ 简化成了 $P(E)$. 另外，John(J)和 Mary(M) 是否打电话直接取决于他们是否听到警铃（A）. 所以可以假设给定 A 时，J 与 B 和 E，以及 M 与 J、B 和 E 都相互独立，即 $P(J\mid B,E,A)=P(J\mid A)$ 和 $P(M\mid B,E,A,J)=P(M\mid A)$. 将这些独立假设放在一起，得

$$P(B,E,A,J,M)=P(B)P(E)P(A\mid B,E)P(J\mid A)P(M\mid A).\tag{2.2}$$

这样就把联合分布 $P(B,E,A,J,M)$ 分解成了若干个复杂度较低的概率分布的乘积. 式（2.2）右端的5个概率分布仅包含 $1+1+4+2+2=10$ 个独立参数，相对于联合分布所需的31个独立参数来说，模型的复杂度得到了降低.

更一般地，考虑一个包含 n 个变量的联合分布 $P(X_1,\cdots,X_n)$. 利用链规则，可以把它写为

$$\begin{aligned}P(X_1,\cdots,X_n)&=P(X_1)P(X_2\mid X_1)\cdots P(X_n\mid X_1,X_2,\cdots,X_{n-1})\\&=\prod_{i=1}^{n}P(X_i\mid X_1,X_2,\cdots,X_{i-1}).\end{aligned}\tag{2.3}$$

对于任意 X_i，如果存在 $\pi(X_i)\subseteq\{X_1,\cdots,X_{i-1}\}$，使得给定 $\pi(X_i)$，X_i 与 $\{X_1,\cdots,X_{i-1}\}$ 中的其它变量条件独立，即

$$P(X_i \mid X_1, \cdots, X_{i-1}) = P(X_i \mid \pi(X_i)), \tag{2.4}$$

那么有

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i)). \tag{2.5}$$

这样，就得到了联合分布的一个分解. 其中当 $\pi(X_i) = \varnothing$ 时，$P(X_i \mid \pi(X_i))$ 为边缘分布 $P(X_i)$.

假设对任意的 X_i，$\pi(X_i)$ 最多包含 m 个变量. 当所有变量均取二值时，式(2.5) 右端所含的独立参数最多为 $n2^m$ 个. 相对于原来确定联合分布所需的 $2^n - 1$ 个参数来说，条件独立使模型得到了简化，这在变量数目 n 很大且 $m \ll n$ 时效果更为显著.

2.3 贝叶斯网的概念

在式(2.5)所示的分解中，变量 X_i 的分布直接依赖于 $\pi(X_i)$ 的取值. 如果给定 $\pi(X_i)$，则 X_i 与 $\{X_1, \cdots, X_{i-1}\}$ 中的其它变量条件独立. Pearl (1986) 提出用如下方法构造一个有向图来表示这些依赖和独立关系：①把每个变量都表示为一个节点；②对于每个节点 X_i，都从 $\pi(X_i)$ 中的每个节点画一条有向边到 X_i.

在例 2.1 中，$\pi(B) = \pi(E) = \varnothing$，$\pi(A) = \{B,E\}$，$\pi(J) = \{A\}$，$\pi(M) = \{A\}$. 按照上述方法，可以得到如图 2.1 所示的有向图. 这个图使得变量之间的依赖和独立关系一目了然：我们看到 A 依赖于 B 和 E；M 和 J 都依赖于 A；而从 B 和 E 没有直接到 M 和 J 的有向边，表示给定 A，这两组变量相互条件独立；等等. 第 3 章将对有向图的图形结构和随机变量的条件独立之间的关系进行深入探讨.

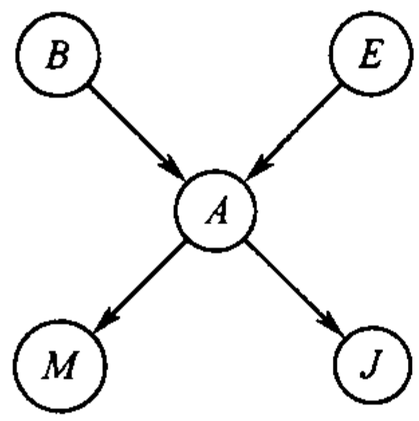
上面提到，变量 A 依赖于 B 和 E，那么 A 具体是如何依赖于 B 和 E 的呢？条件概率分布 $P(A \mid B,E)$ 定量地回答了这个问题：当盗窃和地震都发生时，警铃响的概率 $P(A = y \mid B = y, E = y)$ 是 0.95；当只发生盗窃但没有发生地震时，警铃响的概率 $P(A = y \mid B = y, E = n)$ 是 0.29；当只发生地震但没有发生盗窃时，警铃响的概率 $P(A = y \mid B = n, E = y)$ 是 0.94；而当盗窃和地震都没有发生时，警铃响的概率 $P(A = y \mid B = n, E = n)$ 是 0.001.

类似地，$P(M \mid A)$ 和 $P(J \mid A)$ 分别定量刻画了 M 和 J 如何依赖于 A. 变量 B 和 E 不依赖于其它变量，$P(B)$ 和 $P(E)$ 给出它们的边缘分布. 图 2.1 所示的有向图与这 5 个概率分布合在一起，就构成一个**贝叶斯网** (Bayesian network)，它是对式 (2.2) 给出的联合分布的分解的直观表示.

在给出贝叶斯网的一般定义之前，需要先引入几个图论中的基本概念. 在一个有向图中，如果从节点 X 到节点 Y 有一条边，那么称 X 为 Y 的**父节点** (parent)，而 Y 为 X 的**子节点** (child). 一个节点的所有父节点和子节点称为它的

B	$P(B)$
y	0.01
n	0.99

E	$P(E)$
y	0.02
n	0.98



A	B	E	$P(A\|B,E)$
y	y	y	0.95
n	y	y	0.05
y	y	n	0.94
n	y	n	0.06
y	n	y	0.29
n	n	y	0.71
y	n	n	0.001
n	n	n	0.999

M	A	$P(M\|A)$
y	y	0.9
n	y	0.1
y	n	0.05
n	n	0.95

J	A	$P(J\|A)$
y	y	0.7
n	y	0.3
y	n	0.01
n	n	0.99

图 2.1 Alarm 贝叶斯网：联合分布分解的有向图表达

邻居节点（neighbors）. 没有父节点的节点称为**根节点**（root node），没有子节点的节点称为**叶节点**（leaf node）. 一个节点的**祖先节点**（ancestors）包括其父节点及父节点的祖先节点，根节点无祖先节点. 一个节点的**后代节点**（descendants）包括其子节点及子节点的后代节点，叶节点无后代节点. 一个节点的**非后代节点**（nondescendants）包括所有不是其后代节点的节点. 后面分别记节点 X 的父节 点为 $\mathrm{pa}(X)$ 或 $\pi(X)$，子节点为 $\mathrm{ch}(X)$，邻居节点为 $\mathrm{nb}(X)$，祖先节点为 $\mathrm{an}(X)$，后代节点为 $\mathrm{de}(X)$，非后代节点为 $\mathrm{nd}(X)$. 在图 2.1 中，B，E 是根节点，A 是叶节点 J 的父节点，B，E，A 都是 J 的祖先节点. 在一有向图中，若某节点是它自己的祖先节点，则该图包含一**有向圈**（directed cycle）. **有向无圈图**（directed acyclic graph）是不含有向圈的有向图.

贝叶斯网是一个**有向无圈图**，其中节点代表随机变量[①]，节点间的边代表变量之间的直接依赖关系. 每个节点都附有一个概率分布，根节点 X 所附的是它的边缘分布 $P(X)$，而非根节点 X 所附的是条件概率分布 $P(X \mid \pi(X))$.

贝叶斯网可以从定性和定量两个层面来理解. 在定性层面，它用一个有向无圈图描述了变量之间的依赖和独立关系. 在定量层面，它则用条件概率分布刻画了变量对其父节点的依赖关系. 在语义上，贝叶斯网是联合概率分布的分解的一种表示. 更具体地，假设网络中的变量为 $X_1,\cdots,X_n$，那么把各变量所附的概率分布相乘就得到联合分布，即

$$P(X_1, \cdots, X_i) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i)),$$

① 从现在起，我们对变量和节点不做专门的区分，两者可以混用.

其中当 $\pi(X_i)=\varnothing$ 时，$P(X_i \mid \pi(X_i))$ 即是边缘分布 $P(X_i)$.

联合概率分布的分解降低了概率模型的复杂度. 贝叶斯网的引入虽然没有进一步降低复杂度，但它为概率推理提供了很大的方便. 这主要是因为贝叶斯网一方面是严格的数学语言，适合于计算机的处理；另一方面，它又直观易懂，方便人们讨论交流和建立模型. 另外，Pearl (1986) 还认为，贝叶斯网提供了人脑推理过程的一个模型，因为依赖和独立关系是人们日常推理的基本工具，而且人类知识的基本结构也可以用依赖图来表达. 事实上，越来越多的研究领域开始采用贝叶斯网来展示问题的结构，从而使贝叶斯网的影响远远超出了不确定性推理和人工智能的范围.

2.4 贝叶斯网的构造

贝叶斯网的构造方法有两种，一种是通过咨询专家手工构造，另一种是通过数据分析来获得. 本书第三部分会详细讨论怎样利用机器学习方法分析数据来获得贝叶斯网. 本节则讨论如何手工构造贝叶斯网，这包括确定网络结构和评估条件概率两个子任务. 这些讨论不仅有助于实际应用，还能加深读者对贝叶斯网的理解.

2.4.1 确定网络结构

2.2 节和 2.3 节在引入贝叶斯网概念的同时，也暗示了如下的确定网络结构的方法：

(1) 选定一组刻画问题的随机变量 $\{X_1, X_2, \cdots, X_n\}$；

(2) 选择一个变量顺序 $\alpha=\langle X_1, X_2, \cdots, X_n\rangle$；

(3) 从一个空图出发，按照顺序 α 逐个将变量加入 $\mathscr{G}$ 中；

(4) 在加入变量 X_i 时，$\mathscr{G}$ 中的变量包括 $X_1, X_2, \cdots, X_{i-1}$：

①利用问题的背景知识，在这些变量中选择一个尽可能小的子集 $\pi(X_i)$，使得假设“给定 $\pi(X_i)$，X_i 与 $\mathscr{G}$ 中的其它变量条件独立”合理；

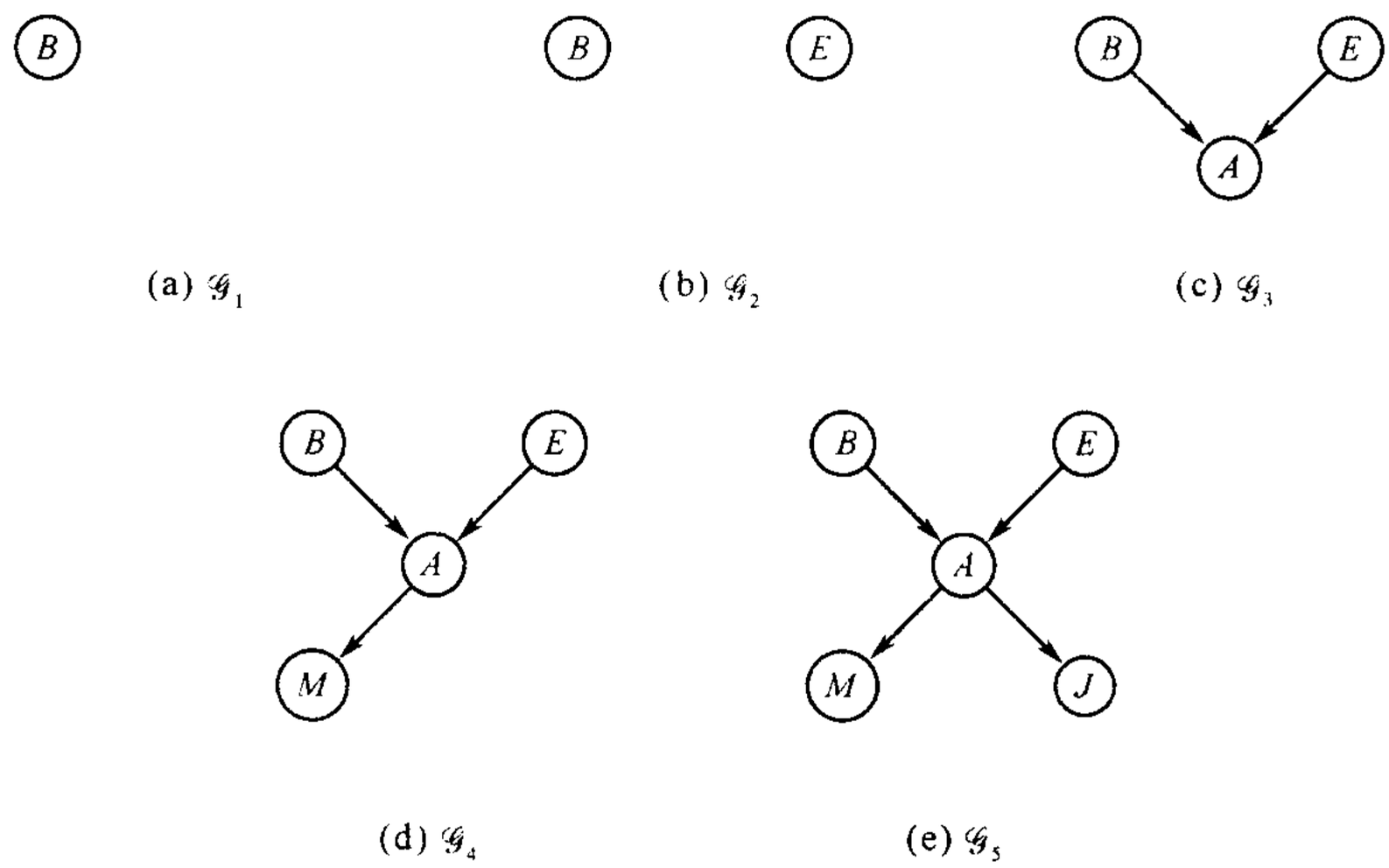
②从 $\pi(X_i)$ 中的每一个节点添加一条指向 X_i 的有向边.

例 2.2 例 2.1 中的 Alarm 问题涉及 5 个随机变量. 假设选用序 $\alpha_1=\langle B, E, A, M, J\rangle$ 来为它构造一个贝叶斯网结构，过程如图 2.2 所示. ①首先把 B 加入空图，得到 $\mathscr{G}_1$；②接着加入 E：假设 B 和 E 相互独立，$\pi(E)=\varnothing$，因此无需加边，结果是 $\mathscr{G}_2$；③然后加入 A：我们认为 A 同时依赖 B 和 E，所以 $\pi(A)=\{B,E\}$，于是分别从 B 和 E 画一条到 A 的边，得到 $\mathscr{G}_3$；④之后加入 M：假设给定 A，M 与 B，E 条件独立，所以 $\pi(M)=\{A\}$，于是从 A 画一条到 M 的边，得到 $\mathscr{G}_4$；⑤最后加入 J：假设给定 A，J 与 B，E 和 M 相互条件独立，所以 $\pi(J)=$

$\{A\}$，于是画一条从 A 到 J 的边，得到 $\mathscr{G}_5$. $\mathscr{G}_1 \sim \mathscr{G}_5$ 分别如图 2.2(a)～(e)所示.

□

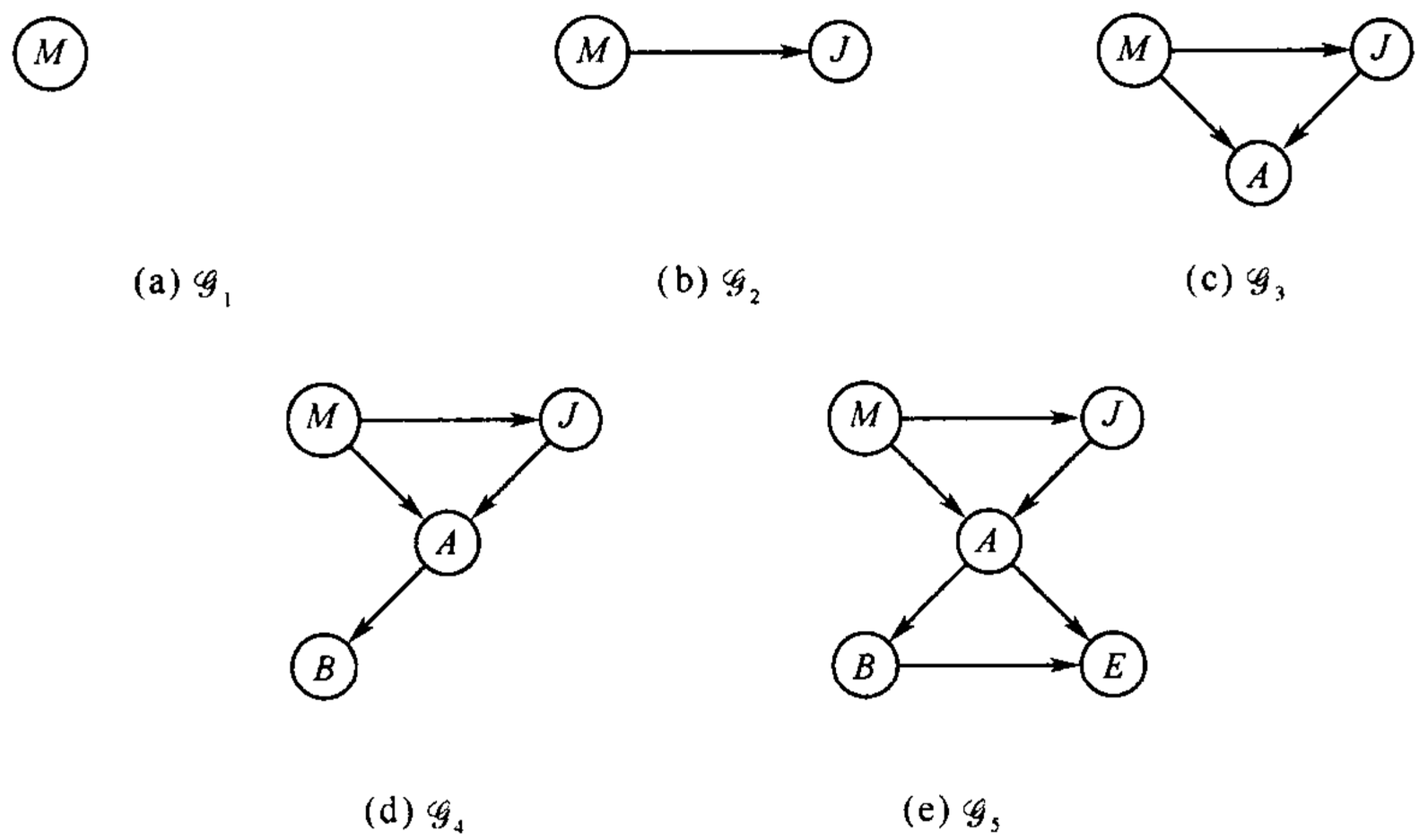
值得注意的是，从不同的变量顺序出发，可能得到不同的网络结构. 下面的两个例子说明了这一点.



(a) $\mathscr{G}_1$　(b) $\mathscr{G}_2$　(c) $\mathscr{G}_3$　(d) $\mathscr{G}_4$　(e) $\mathscr{G}_5$

图 2.2　用序 $\alpha_1 = \langle B, E, A, M, J \rangle$ 构造 Alarm 贝叶斯网结构的过程

例 2.3　在例 2.2 中，如果使用序 $\alpha_2 = \langle M, J, A, B, E \rangle$，则构造贝叶斯网结构的过程如图 2.3 所示. ①首先把 M 加入空图，得到 $\mathscr{G}_1$. ②接着加入 J：在例 2.2 中，我们假设 给定 A,M 和 J 条件独立，但是，在还不知道 A 时，却不能假设 M 和 J 相互独立，因为若 Pearl 教授接到 Mary 的电话，那么他家的警铃响的可能性增大，从而 John 会打电话的可能性也会变大. 因此 M 和 J 相互不独立，从而 $\pi(J) = \{M\}$，于是有边“$M \rightarrow J$”，得到 $\mathscr{G}_2$. ③然后加入 A：是否接到 Mary 和 John 的电话显然都会影响对警铃是否已响的判断，所以 $\pi(A) = \{M, J\}$，于是有边“$M \rightarrow A$”和“$J \rightarrow A$”，得到 $\mathscr{G}_3$. ④接着加入 B：可以假设给定 A,B 与 M 及 J 相互条件独立，所以 $\pi(B) = \{A\}$，于是有边“$A \rightarrow B$”，得到 $\mathscr{G}_4$. ⑤最后加入 E：先考虑它和 A 的关系，E 显然依赖于 A，因为知道警铃响会增加对地震发生的信度. 对 M 和 J，可以假设给定 A，它们与 E 条件独立. 最后考虑 B 和 E 的关系. 在例 2.2 中，我们假定 B 和 E 相互边缘独立，即在不知道其它变量的取值时它们相互独立. 但是，若已知 A 的取值，却不能假设 B 和 E 相互独立. 考虑警铃已响的情况. 若已知有地震发生，则警铃响可能是由地震造成的，从而对发生盗窃的信度会降低；若已知没有地震发生，则排除了地震触响警铃的可能性，于是对发生盗窃的信度会增高. 综上所述，应该假设 $\pi(E) = \{A,B\}$，于是有边“$A \rightarrow E$”和“$B \rightarrow E$”. 最后得到 $\mathscr{G}_5$. $\mathscr{G}_1 \sim \mathscr{G}_5$ 分别如图 2.3



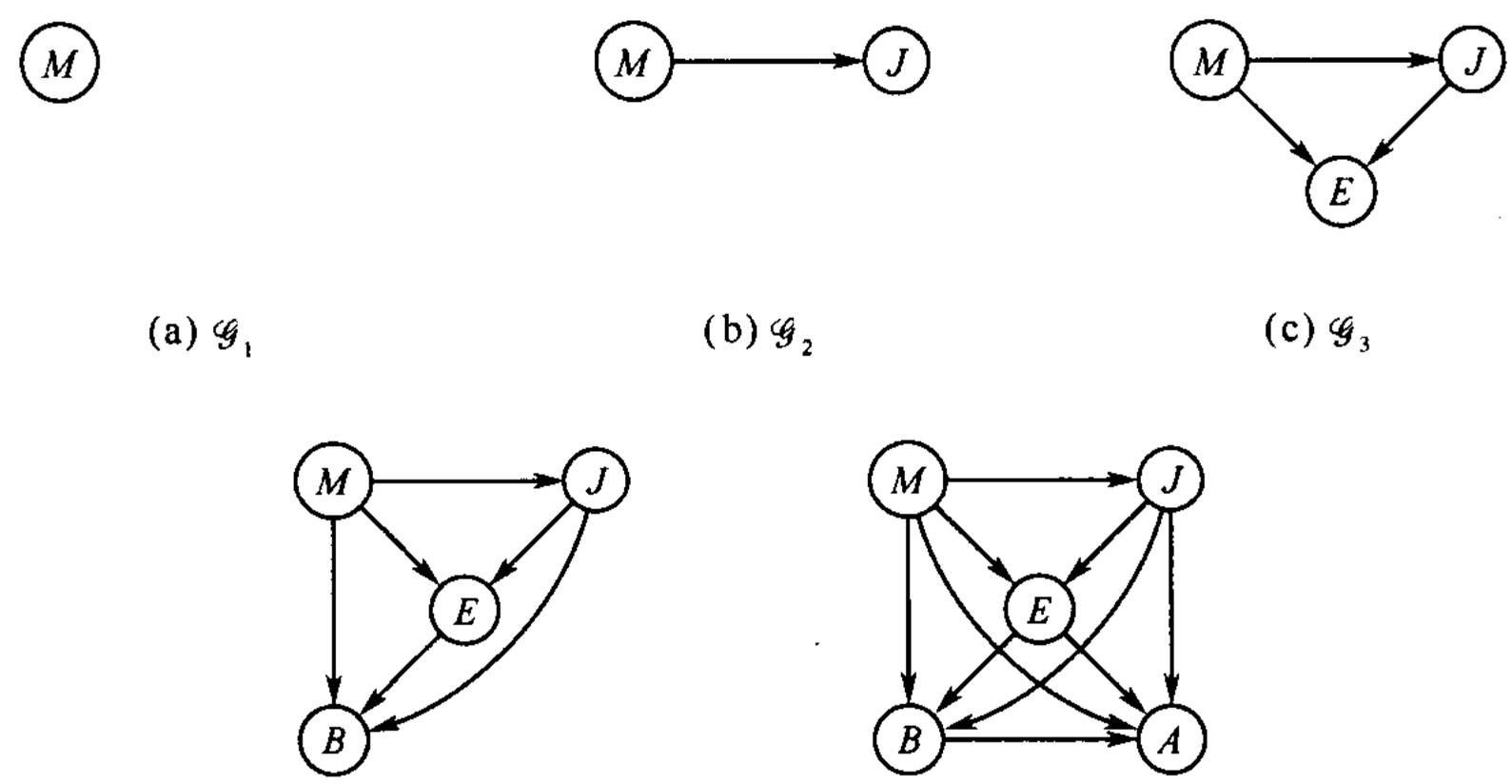
(a) $\mathscr{G}_1$ (b) $\mathscr{G}_2$ (c) $\mathscr{G}_3$

(d) $\mathscr{G}_4$ (e) $\mathscr{G}_5$

图 2.3 用序 $\alpha_2 = \langle M, J, A, B, E \rangle$ 构造 Alarm 贝叶斯网结构的过程

(a)～(e)所示. □

例 2.4 在例 2.2 中，如果使用序 $\alpha_3 = \langle M, J, E, B, A \rangle$，则 Alarm 网络的构造过程如图 2.4 所示. 为节省篇幅，我们把细节留给读者来补充. □



(a) $\mathscr{G}_1$ (b) $\mathscr{G}_2$ (c) $\mathscr{G}_3$

(d) $\mathscr{G}_4$ (e) $\mathscr{G}_5$

图 2.4 用序 $\alpha_3 = \langle M, J, E, B, A \rangle$ 构造 Alarm 贝叶斯网结构的过程

不同的变量顺序导致不同的网络结构，不同的网络结构表示了联合分布的不同分解，而不同的分解则意味着不同的复杂度. 那么，应该按照什么原则来选择变量顺序呢？对这个问题有以下 3 种看法.

（1）Smith（1989）认为应以模型的复杂度为标准. 在例 2.2～2.4 中的 3 个

变量顺序 α_1，α_2 和 α_3 中，应该选择 α_1，因为用它得到的模型最简单.

（2）Howard 和 Matheson（1984）认为变量顺序的选取应以条件概率评估的难易程度为标准. 在 Alarm 例子中，如果选用 α_3，则需确定条件分布 $P(B \mid J, M, E)$，即已知发生地震、也接到了来自 John 和 Mary 的电话时对发生盗窃的概率. 这个分布难以直观评估，因为人们很少会把这几件事全部联想到一起. 而如果选用 α_1，则所需概率分布较易于评估.

（3）Pearl（2000）提出应该用因果关系来决定变量顺序，原因在前，结果在后. 在上面的例子中，α_1 符合因果顺序，而 α_2 和 α_3 则不符合. 实际应用中，因果关系往往使得网络结构简单，概率分布易于评估.

2.4.2　因果关系与贝叶斯网

在实际应用中，人们往往利用因果关系来确定贝叶斯网的结构. 在 Alarm 的例子中，地震（E）和盗窃（B）是警铃响（A）的直接原因，于是有“$E \rightarrow A$”和“$B \rightarrow A$”这两条边；警铃响（A）是 Mary 和 John 打电话的直接原因，于是我们画出边“$A \rightarrow M$”和“$A \rightarrow J$”. 这样，利用因果关系，很容易就获得了图 2.1 中的网络结构. 下面再来看一个例子.

例 2.5　考虑以下变量之间的关系：肺炎（T），肺癌（L），支气管炎（B），出访（A），抽烟（S），呼吸困难（D）及 X 光胸透结果（X）. 我们知道：抽烟会引起肺癌和支气管炎，出访可能感染肺炎，这 3 种疾病的任何一个都可能导致呼吸困难. 另外如有肺炎或肺癌，X 光胸透结果可能会呈阳性. 把这些因果关系综合起来，就得到如图 2.5(a)所示的网络结构. 在图 2.5(b)中，我们引入一个组合变量“肺炎或肺癌（R）”来简化网络结构.　□

在利用因果关系建立起来的贝叶斯网中，变量间的边表示的是因果关系，而

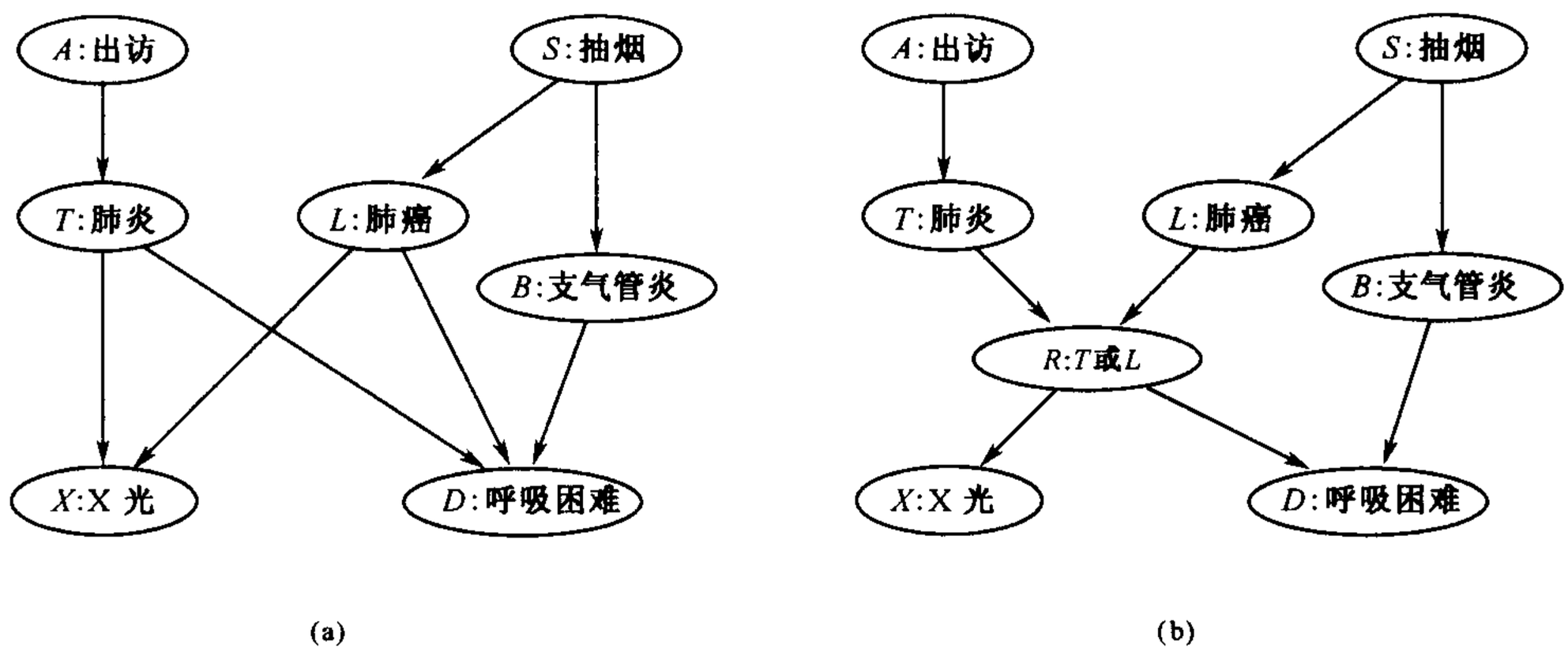
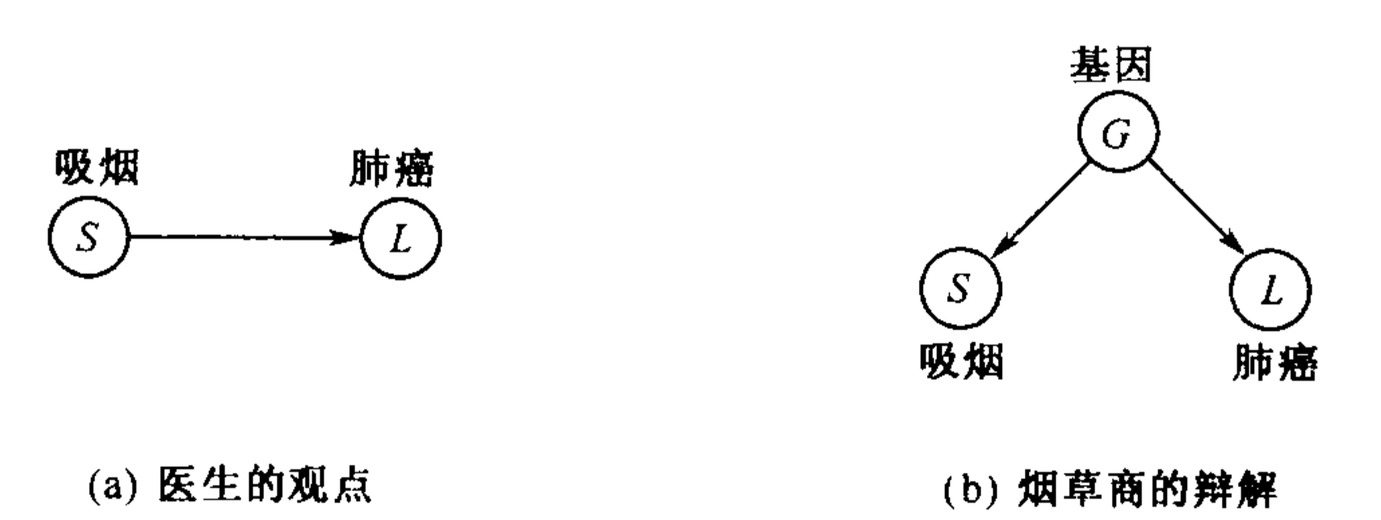


图 2.5　肺病诊断因果网

非简单的概率依赖关联. 这样的贝叶斯网称为**贝氏因果网** (Bayesian causal networks), 简称**因果网** (causal networks). 在贝氏因果网中, 除了可以进行概率推理外, 还可以进行关于**干预后果** (effects of intervention) 的推理以及**虚设** (counter factual) 推理. 本书仅把因果关系当作一种建立贝叶斯网的方法, 而不对因果网做进一步的讨论, 有兴趣的读者可以参见文献(Pearl, 2000).

利用因果关系建立贝叶斯网网络结构有两点需要注意. 首先, 因果关系还没有一个能被广泛接受的严格定义. 对它到底是客观世界本身的属性, 还是人的意识为了理解世界而创造出来的主观概念, 人们还没有达成共识. 有的因果关系十分明显, 例如: 肺癌会导致 X 光胸透结果上有阴影, 支气管炎会导致呼吸困难等; 但有的因果关系却不那么明显, 譬如统计显示在技术领域就业的女性人数明显少于男性, 这是否就意味着性别差异影响到人在技术领域的工作能力? 目前对此还存在争议. 又譬如, 多数医生会认为"抽烟 (S) 导致肺癌 (L)", 但烟草行业却辩解说"存在一个既诱发抽烟 (S) 又能导致肺癌 (L) 的基因 (G)", 这两派观点对应两个不同的因果模型, 如图 2.6 所示. 这两个模型都能解释吸烟与肺癌间的统计关联关系.



(a) 医生的观点　　(b) 烟草商的辩解

图 2.6　吸烟与肺癌的因果关系

在实际应用中, 人们往往采用如下方式来判断因果关系: 假设一个万能的上帝可以介入改变任何变量的状态. 对变量 X 和 Y, 如果知道 X 的状态被上帝改变了会影响你对 Y 的信度, 而反过来知道 Y 的状态被上帝改变并不影响你对 X 的信度, 那么就说 X 是 Y 的原因. 考察地震和警铃响之间的因果关系. 如果假设上帝制造一次地震, 则会认为警铃可能会响; 反过来, 如果知道有人弄响警铃, 则不会认为将发生地震. 因此, 地震是警铃响的原因, 反之不然. 另一个例子是抽烟与手指变黄的关系. 如果强迫某人抽上一段时间的烟, 则他的手指可能会变黄. 但是如果将他的手指染黄, 则不会影响对他是否抽烟的信度, 因此抽烟是手指变黄的原因, 而说明手指变黄不是抽烟的原因.

利用因果关系建立贝叶斯网网络结构还有一点需要注意, 那就是因果关系与条件独立之间的关系. 贝叶斯网蕴含了许多条件独立关系, 这一点在前几节已经

看到，在第 3 章将会看得更加清楚. 所以，当利用因果关系建立贝叶斯网时，实际上是在基于因果关系进行条件独立的假设. 在第 3.2 节将会看到，所做的假设可以归纳为**因果马尔可夫假设**（causal Markov assumption）. 这个假设是因果关系和条件独立之间的桥梁，详细情况请读者参见文献(Spirtes et al.，2000).

2.4.3　确定网络参数

贝叶斯网的参数，即各变量的概率分布，一般是通过数据分析获得的. 不过，有时也可以从问题的特性直接得到.

例 2.6（种马农场）　考虑一个种马农场中的公马、母马和它们所生育的后代之间的基因遗传关系（Jensen，2001）. 假设基因 a 是一个隐性致病基因，对应的显性基因是 A. 当没有任何信息时，任意一匹马关于该疾病的基因型可能是以下三者之一：aa（患病），aA（携带者）或 AA（正常）. 根据基因遗传学，可以直接确定任意一匹马的基因型 G_C 与它的父母的基因型 G_F 和 G_M 之间的概率关系 $P(G_C \mid G_F, G_M)$，如表 2.2 所示. 考查表的第一行，当公马的基因是 aa 时，如果母马也是 aa，则根据基因遗传交换的规律，后代只可能是 aa，所以其分布是 (1,0,0)；如果母马是 aA，则后代不可能是 AA，而有 50%的可能是 aa，50%的可能是 aA，因此其分布是 (0.5,0.5,0)；如果母马是 AA，则后代一定是 aA，因此其分布是 (0,1,0). 表中其余概率分布之意义可依次类推. □

表 2.2　种马基因遗传的概率分布 $P(G_C \mid G_F, G_M)$

G_F \ G_M	aa	aA	AA
aa	(1, 0, 0)	(0.5, 0.5, 0)	(0, 1, 0)
aA	(0.5, 0.5, 0)	(0.25, 0.5, 0.25)	(0, 0.5, 0.5)
AA	(0, 1, 0)	(0, 0.5, 0.5)	(0, 0, 1)

注：数字 (α, β, γ) 分别表示后代基因型是 (aa, aA, AA) 的概率.

例 2.7（信道编码）　假设用一个噪音信道传递信息位 U，信道的出错概率为 0.1. 为提高传输准确度，可以简单地把 U 重复传输 3 次，然后在接收方根据接收结果进行解码，推测正确的 U 的取值. 这个过程可以用如图 2.7中所示的贝叶斯网来表示。网络中所有变量均取二值，其中传输位 X_1，X_2，X_3 是 U 的简单重复，接收位 Y_1，Y_2，Y_3 是接收方接

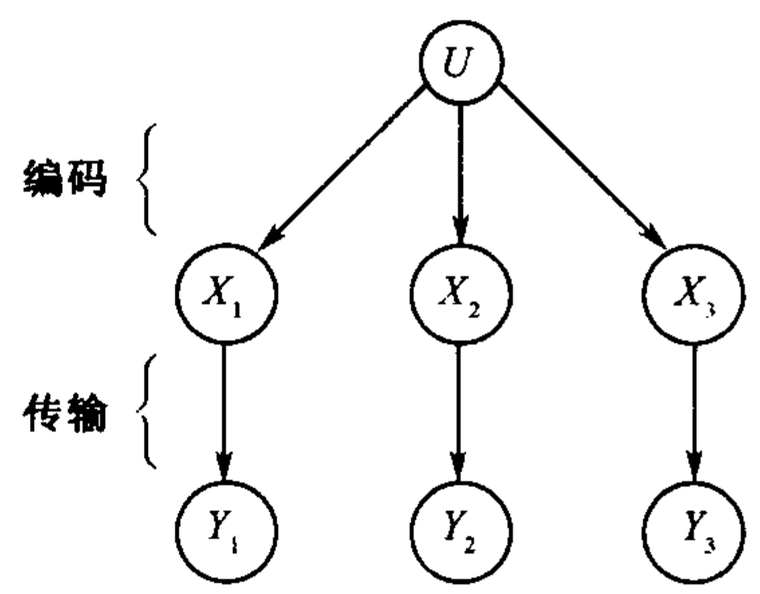


图 2.7　纠错码贝叶斯网示例

收到的 3 个信息位. 可以根据问题的特性来确定网络参数：对于信息位 U 和传输位 X_1，X_2，X_3，有

$$P(U) = (0.5, 0.5);$$

$$P(X_i \mid U) = \begin{cases} 1, & \text{如果 } U \text{ 和 } X_i \text{ 状态相同;} \\ 0, & \text{如果 } U \text{ 和 } X_i \text{ 状态不同;} \end{cases}$$

而接收位 Y_i 的条件概率分布由噪音信道的特性决定：

$$P(Y_i \mid X_i) = \begin{cases} 0.9, & \text{如果 } Y_i \text{ 和 } X_i \text{ 状态相同;} \\ 0.1, & \text{如果 } Y_i \text{ 和 } X_i \text{ 状态不同.} \end{cases}$$

□

有时候，网络参数是通过专家咨询得到的，这在贝叶斯网发展的早期尤为常见. 专家咨询耗费人力和时间，因此有必要尽量减少参数的个数，通常这也有利于提高咨询结果的质量. 下面介绍两个减少参数的方法.

设变量 Y 有 m 个父节点 $X_1, \cdots, X_m$，条件分布 $P(Y \mid X_1, \cdots, X_m)$ 刻画 Y 对其父节点的依赖关系，如图 2.8 所示. 当所有变量均取二值时，它包含 2^m 个独立参数. 为减少参数个数，人们往往假设条件分布具有某种规律，称为局部结构①. 常见的局部结构有两种：**因果机制独立**（causal independence）（Heckerman，1993；Zhang and Poole，1996）和**环境独立**（context specific independence）（Boutilier et al.，1996）.

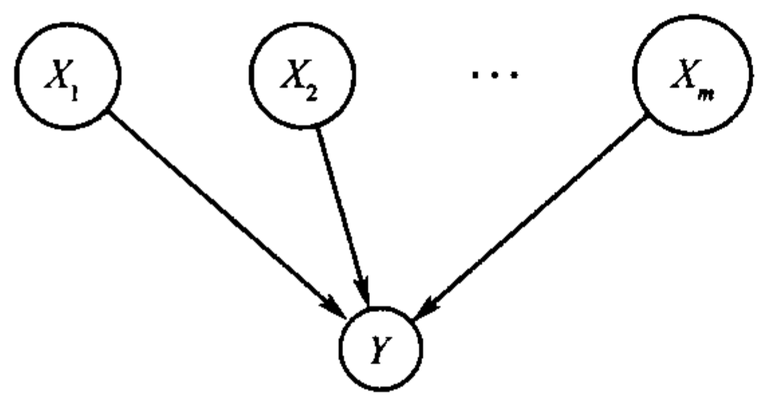


图 2.8 变量 Y 及其 m 个父节点

1. 因果机制独立

因果机制独立指的是多个原因独立地影响同一个结果. 在例 2.1 中，地震（E）和盗窃（B）都可以触发警铃（A），但是两者的机制不同，因此可以假设地震和盗窃独立地影响警铃响，如图 2.9（a）所示. 为了把它的含义讲得更准确些，引入两个辅助变量 A_e 和 A_b，A_e 代表警铃因地震而响，A_b 代表警铃因盗窃而响，故 $A = A_e \vee A_b$. 因果机制独立假设的确切含义是：变量 A_e 和 A_b 相互独立. 上述各变量的关系如图 2.9(b)所示. 这里 E，B，A 之间是一个**噪音或**（noisy OR）的关系（Good，1961）.

设图 2.8 中的箭头表示因果关系. 我们说 $X_1, \cdots, X_m$ 独立地影响 Y，如果存在与 Y 有共同状态空间的变量 $\xi_1, \cdots, \xi_m$，使得：

（1）对每个 i，ξ_i 依赖于 X_i，并且给定 X_i，ξ_i 独立于其它的 ξ_j 和 X_j；

① 这是相对于贝叶斯网结构而言的. 贝叶斯网展现的是联合概率的结构，称为整体结构.

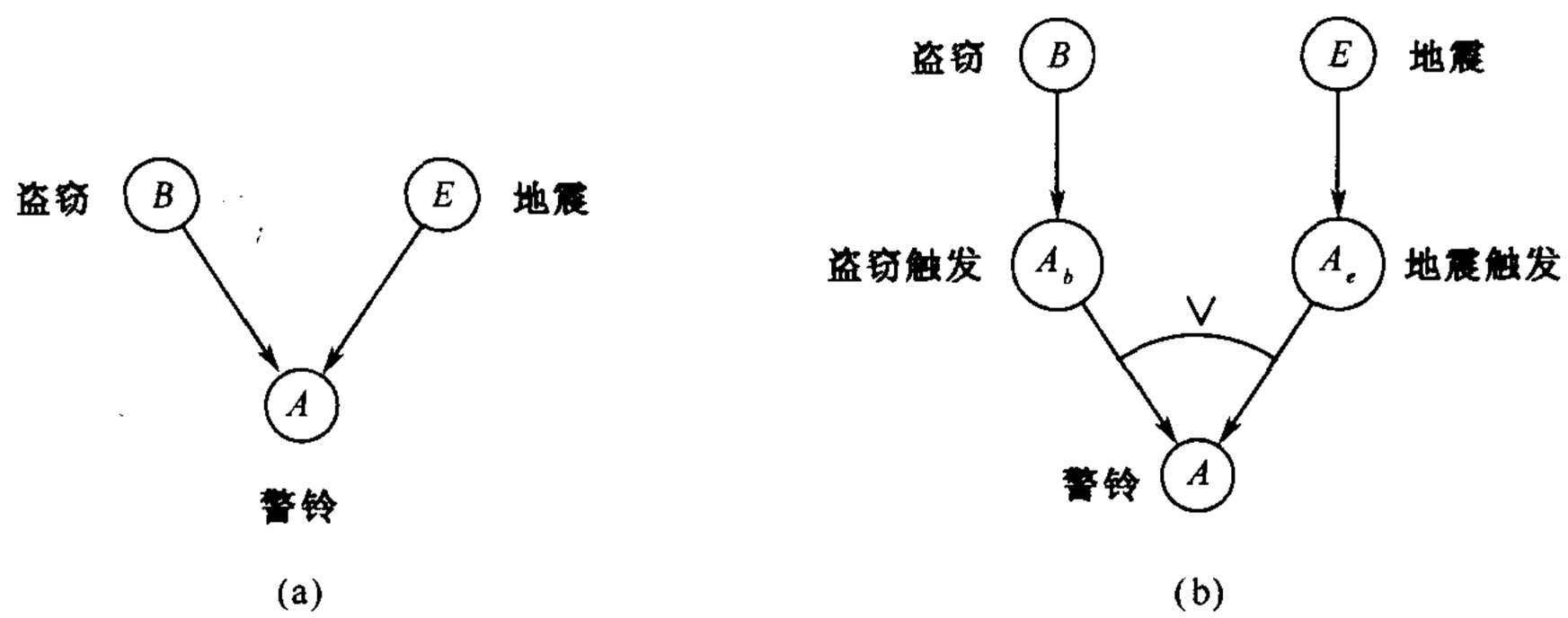


图 2.9 一个因果机制独立的例子

(2) 在 Ω_Y 上存在一个满足交换律和结合律的算子 $*$，使得

$$Y = \xi_1 * \xi_2 * \cdots * \xi_m.$$

这就是说，不同原因对 Y 的影响是独立的，总影响是各原因的单独影响按算子"$*$"的合成结果，如图 2.10 所示. 以下，我们把 ξ_i 称为 X_i 对 Y 的**贡献**，把"$*$"称为**基本合成算子**，把 $P(\xi_i \mid X_i)$ 称为 X_i 对 Y 的**贡献概率分布**.

在变量取二值的情况下，当基本合成算子"$*$"是逻辑或"$\vee$"时，图 2.10 所示的是**噪音或门**（noisy OR gate），当它是逻辑与"$\wedge$"时，图 2.10 所示的是**噪音与门**（noisy AND gate）. **噪音最大（小）门**（noisy MAX/MIN gate）是它们的自然推广，此时"$*$"是最大（小）值运算. 当"$*$"是加法运算时，图 2.10 所示的模型称为**噪音加法器**（noisy adder）.

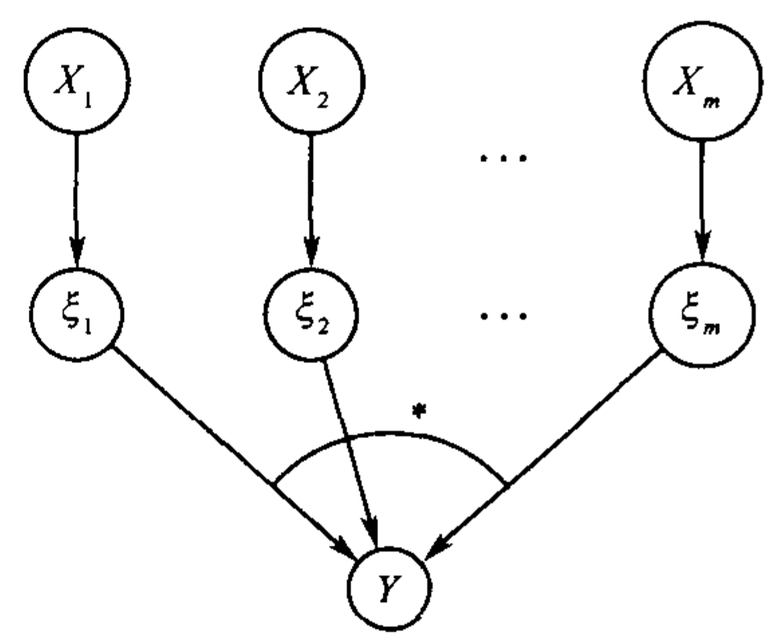


图 2.10 因果机制独立示意图

定理 2.1 设原因变量 $X_1, \cdots, X_m$ 独立地影响结果变量 Y，那么对任意 $\alpha \in \Omega_Y$，有

$$P(Y = \alpha \mid X_1, \cdots, X_m) = \sum_{\alpha_1 * \cdots * \alpha_m = \alpha} P(\xi_1 = \alpha_1 \mid X_1) \cdots P(\xi_m = \alpha_m \mid X_m). \tag{2.6}$$

其中的"$*$"是基本合成算子. □

此定理的证明参见文献（Zhang and Poole，1996）. 它说的是条件概率分布 $P(Y = \alpha \mid X_1, \cdots, X_m)$ 可以从各原因的贡献概率分布出发得到，这大大减少了参数的个数. 在所有变量均取二值时，式（2.6）右端的概率分布所包含独立参数的个数是 $2m$ 个，而不是 2^m 个.

2. 环境独立

环境独立是指在特定环境下才成立的条件独立关系. 一个**环境**（context）是一组变量及其取值的组合. 在例 2.5 中，$\{S=y\}$ 是一个环境，$\{A=y,X=n,S=y\}$ 也是一个环境. 设一个环境所涉及的变量的集合是 **C**，其取值是 **c**，我们用 **C=c** 来记该环境.

设 **X**，**Y**，**Z**，**C** 是 4 个两两交空的变量集合，C 的取值是 **c**. 我们说**在环境 C=c 中 X 与 Y 在给定 Z 时相互条件独立**，如果当 $P(\mathbf{Z},\mathbf{Y},\mathbf{C}=\mathbf{c})>0$ 时，有

$$P(\mathbf{X}\mid\mathbf{Z},\mathbf{Y},\mathbf{C}=\mathbf{c})=P(\mathbf{X}\mid\mathbf{Z},\mathbf{C}=\mathrm{c}). \tag{2.7}$$

若 **Z** 为空，则称 **X** 与 **Y** **在环境 C=c 中相互独立**.

图 2.11 给出了两个环境独立的例子. 在图 2.11（a）中，母牛的产犊总数依赖于年龄，但公牛无论年龄多大，产犊数总是 0. 所以在“$G=$公牛”这个环境中，变量 N 与变量 A 总是相互独立. 由于这个环境独立关系，$P(N\mid A,G)$ 所包含的参数个数从 $2\mid A\mid(\mid N\mid-1)$ 降至 $(\mid A\mid+1)(\mid N\mid-1)$.

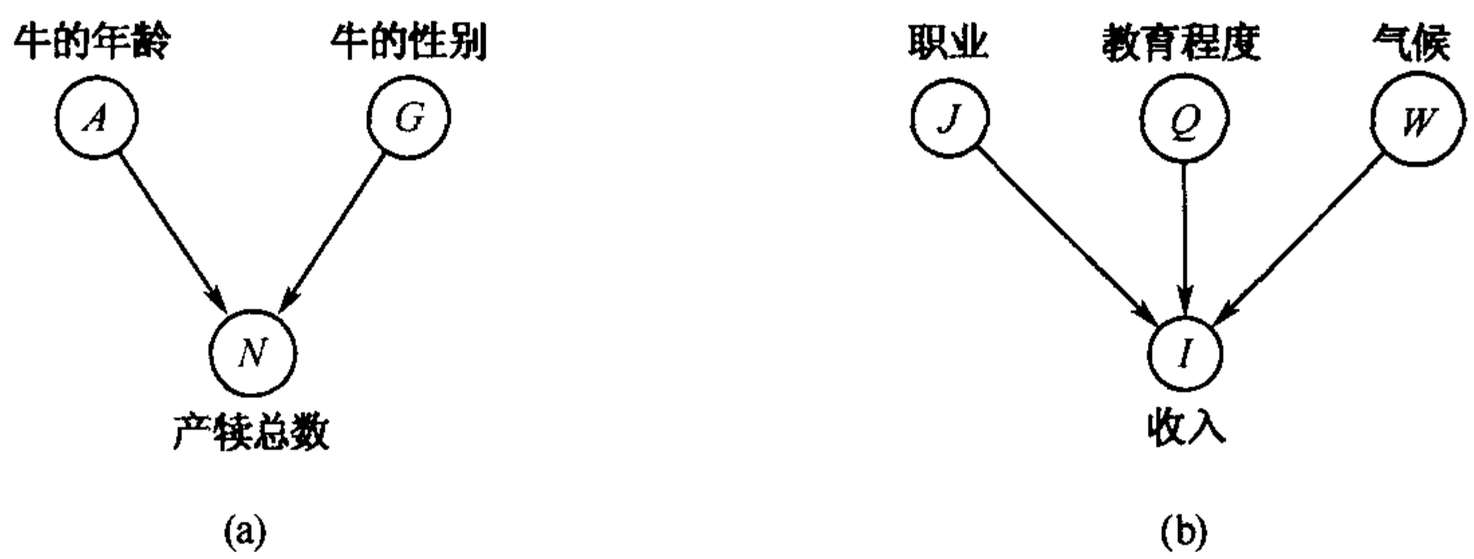


图 2.11 环境独立的例子

在图2.11(b) 中，一个人的收入(I)一般来讲依赖于其职业(J)、受教育程度(Q)以及气候(W). 假设对于一个程序员来说，他的收入不依赖于气候，即在“$J=$程序员”这个环境中，I 与 W 相互独立，有

$$P(I\mid J=\text{程序员},Q,W)=P(I\mid J=\text{程序员},Q).$$

假设对于一个农民来说，他的收入不依赖于受教育程度，即在“$J=$农民”这个环境中，I 与 Q 相互独立，有

$$P(I\mid J=\text{农民},Q,W)=P(I\mid J=\text{农民},W).$$

上述环境独立关系使得参数减少了 $(\mid W\mid-1)\mid Q\mid(\mid I\mid-1)+(\mid Q\mid-1)\cdot\mid W\mid(\mid I\mid-1)$ 个.

2.5　贝叶斯网的应用

这一节和下一节分别介绍贝叶斯网在实际中的应用以及它对其它研究领域的影响. 我们将提供丰富的参考文献，以方便读者深入了解所感兴趣的例子. 在第 10 章中，我们还会详细介绍贝叶斯网在中医研究中的一个应用实例.

2.5.1　医疗诊断

医疗诊断是从一系列临床观测和化验结果出发，对病人所患疾病的类别及其程度进行判断. 在贝叶斯网发展早期，人们研究开发了多个规模可观的医疗诊断网络，比较著名的包括 PATHFINDER（Heckerman，1988，1990）、MUNIN（Andreassen et al.，1989）、CHILD（Spiegelhalter et al.，1993）、QMR-DT 和 CPCS（Shwe et al.，1991；Parker and Miller，1987）等. 下面简单介绍一下这几个网络.

1. PATHFINDER 网络

在医疗诊断中，通常先由病理学家（pathologist）对患者的患病组织切片进行研究诊断，然后再由医师根据病理学家的诊断结果制定合适的治疗方案. 在实际中，一个病理学家经常需要处理许多不同的组织切片，有些不一定是他熟悉的领域，因此错误时有发生. 尤其是在淋巴结组织的诊断中，缺乏经验的年轻病理学家们最容易出错. 为帮助他们提高诊断水平，研究人员开发了一个叫做 PATHFINDER 的专家系统（Heckerman，1988，1990）.

PATHFINDER 先后有 4 个版本，反映了 20 世纪 80 年代人工智能研究者对处理不确定性方法的各种探索. PATHFINDER I 采用了基于规则的系统，不考虑不确定性. PATHFINDER II 试验了包括确定因子（certainty factor）、Dempster-Shafer 理论和朴素贝叶斯模型（naïve Bayes model）在内的多种不确定推理方法，结果显示朴素贝叶斯模型优于其它方法. PATHFINDER III 重新评审了朴素贝叶斯模型中的所有概率，特别关注了那些概率虽小但仍可能发生的事件. PATHFINDER IV 则使用一个更一般的贝叶斯网，能表达朴素贝叶斯模型无法表达的变量之间的依赖关系. 这个贝叶斯网是由开发人员和医疗专家一起手工建造的，确定变量及其取值用了 8 个小时，确定网络拓扑结构用了 35 个小时，对总共 14000 个条件概率值的评估又用了 40 个小时，最终得到了一个有 121 个节点、195 条边的网络. 图 2.12 显示了其网络结构的一部分，网络中的“疾病”节点有 63 个取值，代表淋巴结的 63 种不同疾病，其它节点代表病人的症状，疾病节点是所有症状节点的父节点.

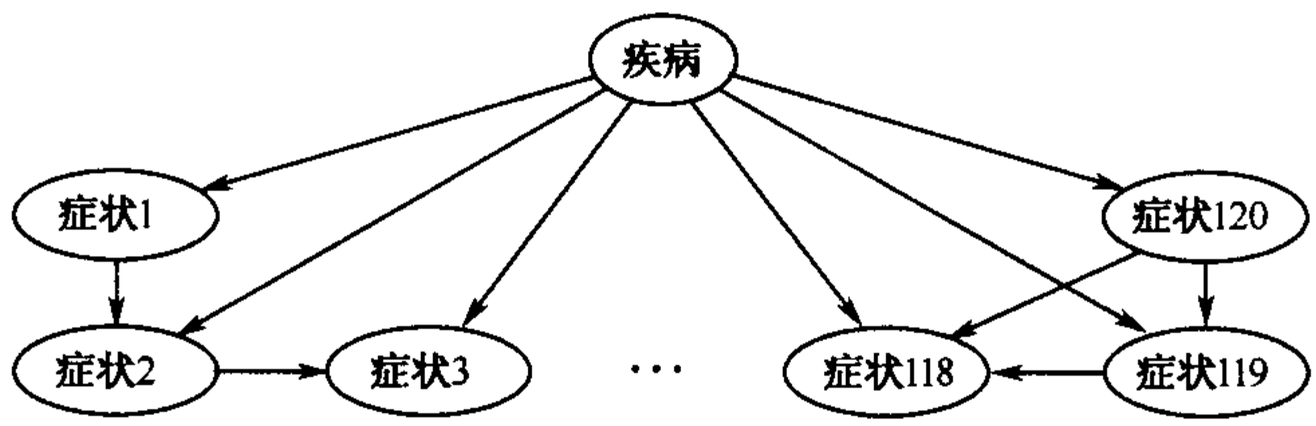


图 2.12　PATHFINDER 网络：疾病节点是所有症状节点的父节点

用 PATHFINDER III 和 IV 进行诊断都是贝叶斯网推理问题，即给定症状变量（findings）的取值，计算疾病变量（disease）的后验分布，P(disease | findings)，然后把后验概率最大的那个疾病作为诊断结果. 系统测试使用了 53 个淋巴腺疾病的病例. 结果显示，贝叶斯网的使用使得诊断水平相对于使用朴素贝叶斯模型时得到了明显提高.

2. CHILD 网络

伦敦 Ormond 街儿童医院（GOS）是一个治疗新生儿先天性心脏病的专业中心，每年治疗大约 200 个来自英国东南部的案例. 新生婴儿如果发现皮肤发紫或有心衰竭征兆，如无法呼吸等，就可能有先天性心脏病，需要及时送往专业中心治疗. 由于病情可能很快恶化，运送之前就需要采取一些措施，这要求专业中心通过电话了解婴儿的情况，并迅速做出诊断. GOS 专业中心 24 小时运作，经常有年轻医生值班，院方希望能用一个计算机系统来改进诊断水平，同时给年轻医生提供训练的机会. CHILD 网络就是在这样的背景下产生的.

CHILD 网（Spiegelhalter et al.，1993）的网络结构如图 2.13 所示，它包含 20 个变量和 25 条边. 网络中，变量 n_2 有 6 个可能的取值，对应 6 种不同的疾病，其它变量平均有 3 个可能的取值. 定量信息方面，它有 114 个概率分布，230 个概率值. 它的网络结构先是通过专家咨询获得，然后再用数据对其进行了修正.

系统测试使用了 168 个病例. 在用数据修正其参数前，CHILD 网能正确诊断其中的 99 例. 在参数修正后，它正确地诊断了 110 例. 相比之下，GOS 的专家能正确地诊断 114 例，其它医院的医师正确诊断了 76 例. 总的结论是，CHILD 网的诊断准确率达到了 GOS 中级儿科专家的水平.

3. QMR-DT 和 CPCS 网络

在 20 世纪 70 年代，匹兹堡大学开发了一个用于内科疾病诊断和医疗教学的

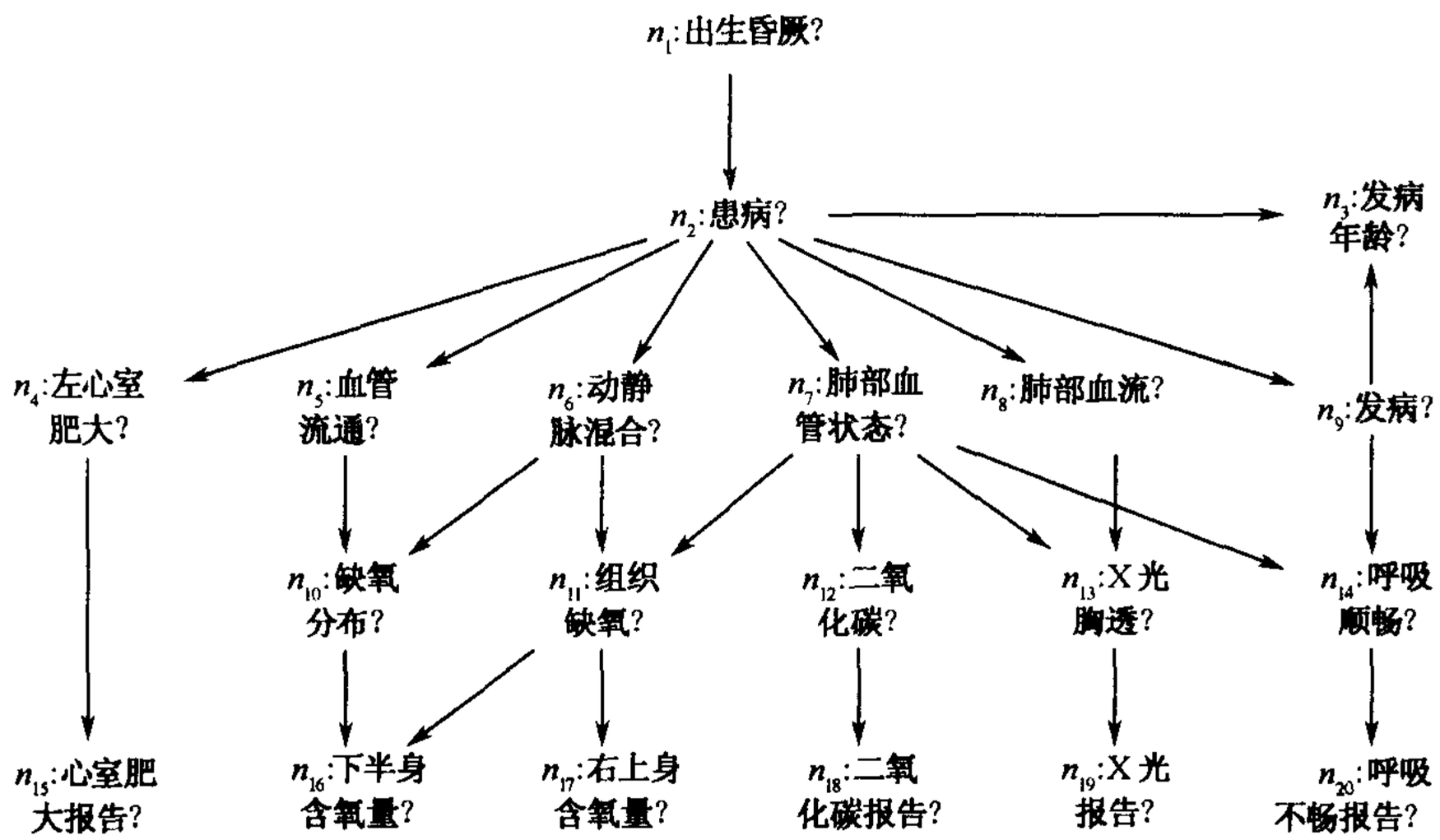


图 2.13　CHILD 网络：用于诊断可能导致“蓝色婴儿”的先天性心脏病

专家系统 INTERNIST-1，它是个基于规则的产生式系统（Miller et al.，1982）. 到了 80 年代，在 INTERNIST-1 知识库的基础上，研究人员又开发了一个商业化的内科疾病诊断的决策支持系统，QMR（quick medical reference）（Miller et al.，1986），它包括 600 多种内科疾病和约 5000 种临床病症. 随着后来概率方法在专家系统中的兴起，人们尝试把基于规则的 QMR 系统转变为基于概率的系统，就得到 QMR-DT 贝叶斯网（Shwe et al.，1991）. 它是一个双层的“噪音或”网络，上层包括 500 多个疾病节点，下层则包括约 4000 个症状节点，所有节点均取二值，上下两层节点之间共有 40000 多条边，如图 2.14 所示.

另一个大型的医疗仿真贝叶斯网 CPCS（computer-based patient case simu-

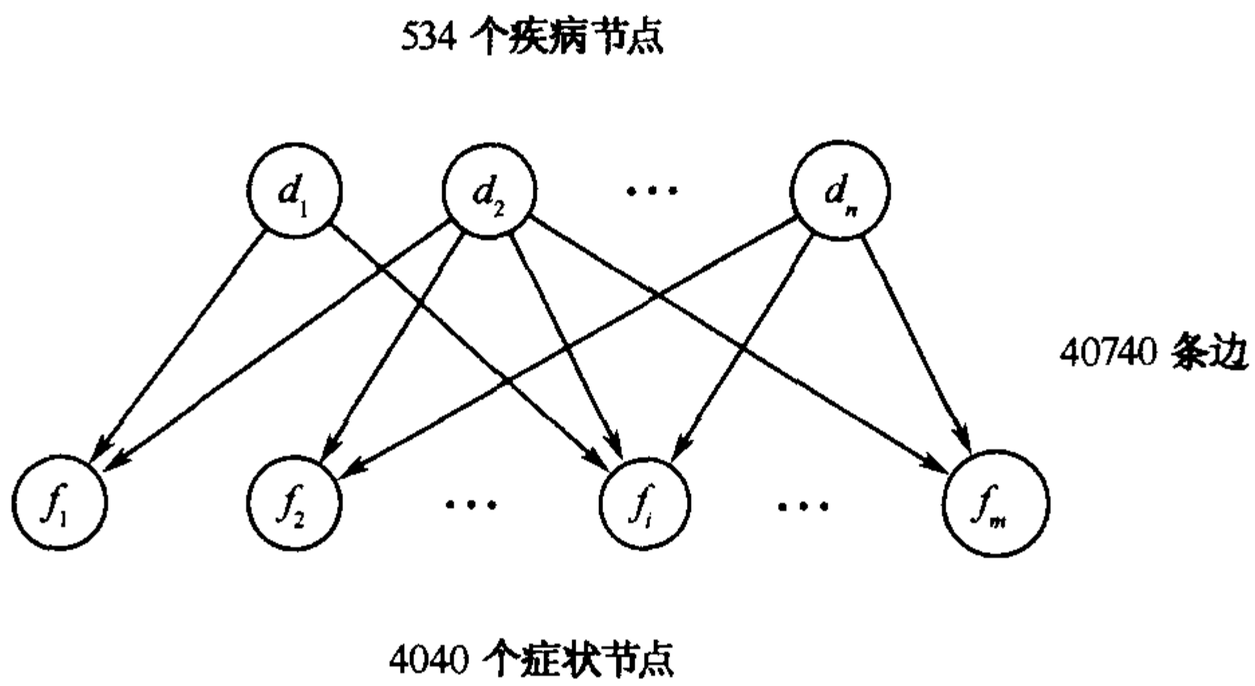


图 2.14　QMR-DT 的网络结构：上层为疾病节点，下层为症状节点

lation)，是从 INTERNIST-1 扩展得到的（Parker and Miller，1987）. 它是一个多层贝叶斯网，共有 448 个节点，908 条边. QMR-DT 和 CPCS 这两个大型网络为开发更高效率的推理算法提出了挑战，它们经常被用来测试评估新开发的推理算法的性能.

2.5.2 工业应用

贝叶斯网在工业中的应用很广，涉及金融分析（Abramson，1991，1994）、产品设计（Corney，2000）、生产制作工艺（Gebhatdr et al.，2003）、工业过程监控管理（Wolbrecht et al.，2000；Weidl et al.，2003a，b）、在线故障诊断（Morjaia et al.，1993；Lerner et al.，2000；Liu and Zhang，2003）以及可靠性分析（Torres-Toledano and Sucar，1998；Langseth，2002；Portinale and Bobbio，1999）等. Friis-Hansen（2001）专门研究了贝叶斯网在海洋工业（maritime industry）中的应用.

下面简单介绍贝叶斯网在故障诊断（fault diagnosis）中的应用. 故障诊断的目的是找出导致一个控制系统失灵的故障部件. 自动故障诊断往往需要在系统中嵌入各种传感器，以监视系统各部件的运行状态，并通过实时推理，及时发现出故障的部件. 在汽车或飞机的关键部件安置传感器，可以预防可能发生的意外；制造系统的故障诊断对工业界也很重要.

基于规则的系统曾被用于自动故障诊断，但它不能反映实际环境中的各种不确定性. 相比之下，贝叶斯网能更好地描述故障诊断中所涉及的不确定性. 图 2.15所示是一个用于汽车启动故障诊断的贝叶斯网（Heckerman et al.，1995）. 从图中可以看到，可能导致汽车无法启动的原因有多个，故障诊断就是根据观测到的证据进行概率推理，找出后验概率最大的那个原因.

其它故障诊断的例子还包括：在线设备维护（D'Ambrosio，1993；D'Ambrosio and Burgess，1996）、商用波音飞机的故障诊断（Kipersztok and Provan，2003）、航天飞机推动引擎的故障诊断（Liu and Zhang，2003；Morris and Beling，2001）以及核电厂中软硬件系统的故障诊断（Santoso et al.，1999；Yildiz et al.，2001；Atte，2001）等.

2.5.3 金融分析

在金融分析（financial analysis）中，贝叶斯网被用于解决石油价格预测（Abramson，1991；Abramson and Finizza，1991；Abramson，1994）、证券风险与回报（portfolio risk and return）（Shenoy and Shenoy，1999）、风险投资决策（venture capital decision）（Kemmerer et al.，2002）以及运筹风险分析（operational risk analysis）（Neil and Tranham，2002）等问题.

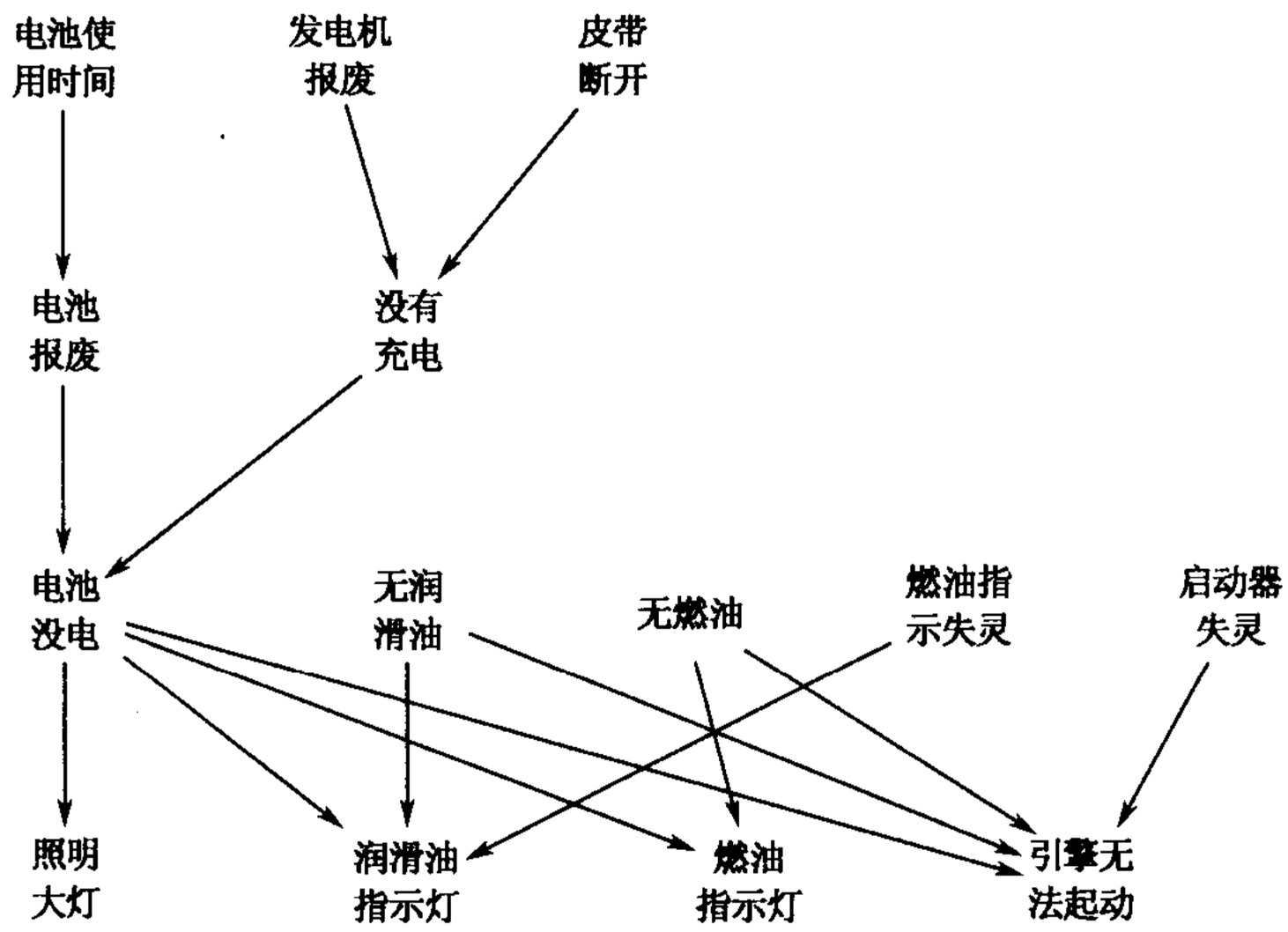


图 2.15　用于汽车启动故障诊断的贝叶斯网

下面简单介绍一下文献(Shenoy and Shenoy，1999)中利用贝叶斯网所做的证券风险与回报分析研究. 传统的证券风险与回报分析主要是基于描述经济变量之间的定量关系的金融模型进行分析和预测. 但在实际中，往往还需要考虑金融专家关于某些信息对股票回报、利率等的影响的评估意见. 例如，对微软不信任案的判决结果会影响到很多公司的股票. 这些主观评估意见与历史数据的结合，能够更为准确地对相关证券的风险与回报进行预测. 但是，这类信息很难体现于传统的金融模型中. 而贝叶斯网则为传统金融分析模型的定量内容和专家主观的定性判断的结合提供了一条合适的途径，使得传统模型在得到新信息时能够系统地更新相应的回报分布情况. 在文献(Shenoy and Shenoy，1999) 中，传统模型先被表示为一个贝叶斯网，然后在其基础上引入专家的主观知识. 在得到网络结构后，定量的信息再被加入. 基于这个贝叶斯网进行分析的结果是一个证券回报概率分布，进一步还可以计算回报的期望值、期望方差及风险等信息. 图 2.16 所示的是一个简单的用于证券风险回报分析的贝叶斯网，其中 ABX、AEM 和 BGO 是 3 家从事黄金开采的公司，它们各自的股票回报被表示为网络中的 3 个同名变量，它们都依赖于股票市场、黄金价格以及与各自股票相关的一个效应因子.

2.5.4　计算机系统

贝叶斯网在计算机系统中的应用包括程序理解（program understanding）(Breese and Blake，1995；Burnell and Horvitz，1995)、软件测试（software

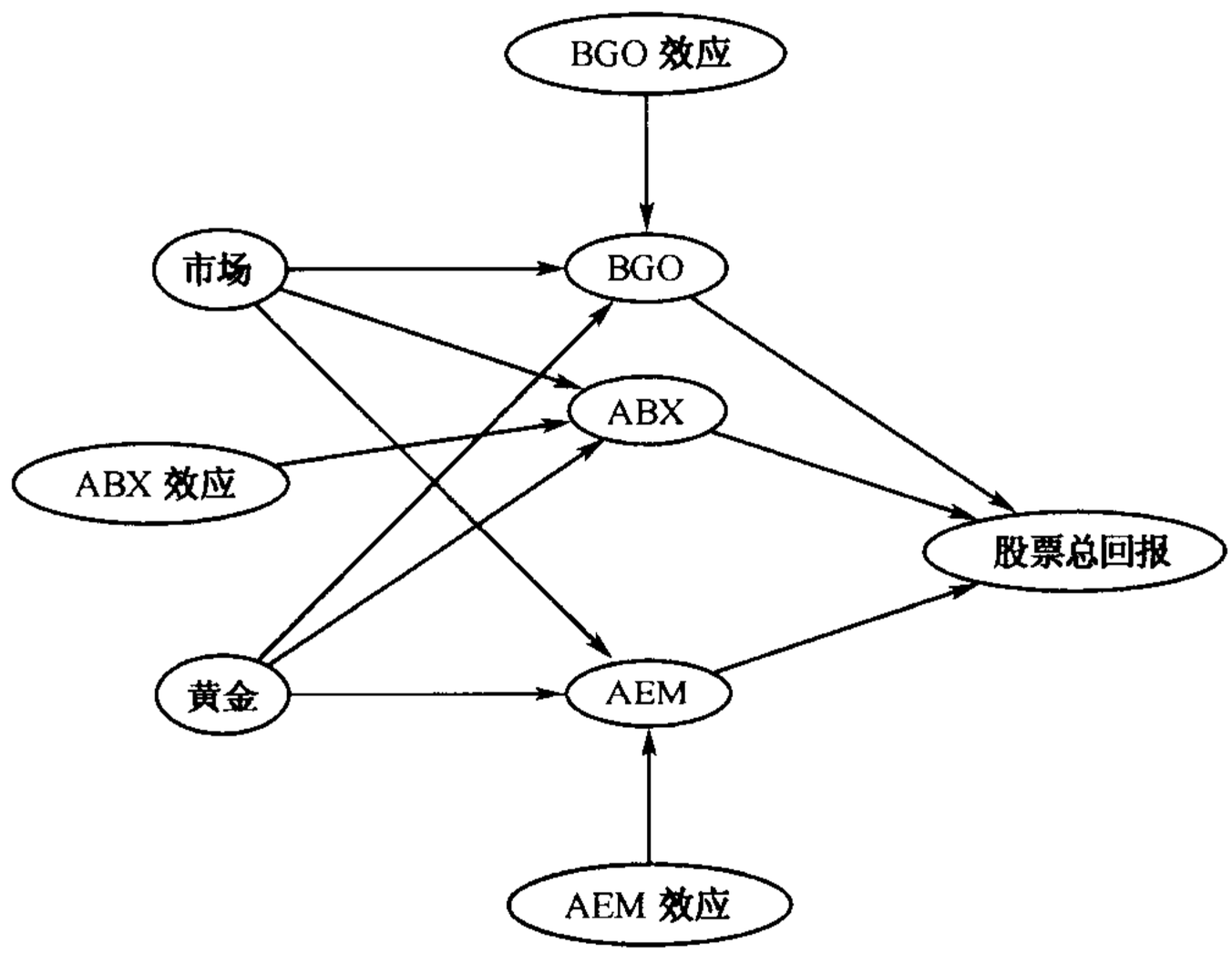


图 2.16 用于证券回报预测的贝叶斯网

testing）（Ziv and Richardson，1997）、垃圾邮件过滤（junk e-mail filtering）（Sahami et al.，1998）、计算机系统故障诊断（trouble shooting）（Heckerman et al.，1995；Jensen et al.，2001）、决策系统信息显示（information display）（Horvitz and Barry，1995）、信息提取（information retrieval）（Fung and Favero，1995；Wong and Butz，2000；Ruokangas and Mengshoel，2003；Blei et al.，2003）、用户特征提取（user profiling）（Horvitz et al.，1998；Schiaffino and Amandi，2000）等. 微软研究院在这方面投入了相当大的力量，开发了一系列嵌入 Windows 和 Office 等系统的贝叶斯网，如 Windows 中的打印机故障诊断程序，Office 中的用户助手，以及 Outlook 中的垃圾邮件过滤器等.

下面简单介绍一个用于打印机故障诊断的贝叶斯网，图 2.17 给出了它的一个局部. 当出现打印颜色过浅的问题时，可能的故障有多个，包括墨粉不均匀、墨盒故障、数据出错和打印驱动故障等. 这些故障可用如下措施修复：取出墨盒摇晃重装、换墨盒、关掉电源重启等. 另外还可以向用户提问，收集相关信息，以帮助进一步查找原因，各变量具体的条件分布表见文献（Vomlel，2003），更复杂的打印机故障诊断网络见文献（Heckerman et al.，1995；Jensen et al.，2001）.

2.5.5 军事应用

战场上局势复杂多变，充满不确定性，其中所涉及的推理问题往往具有实时性、动态性以及离散和连续变量相混合等特点. 贝叶斯网在军事上的应用包括目标识别（target identification）（Hautaniemi et al.，2000）、多目标跟踪（multi-

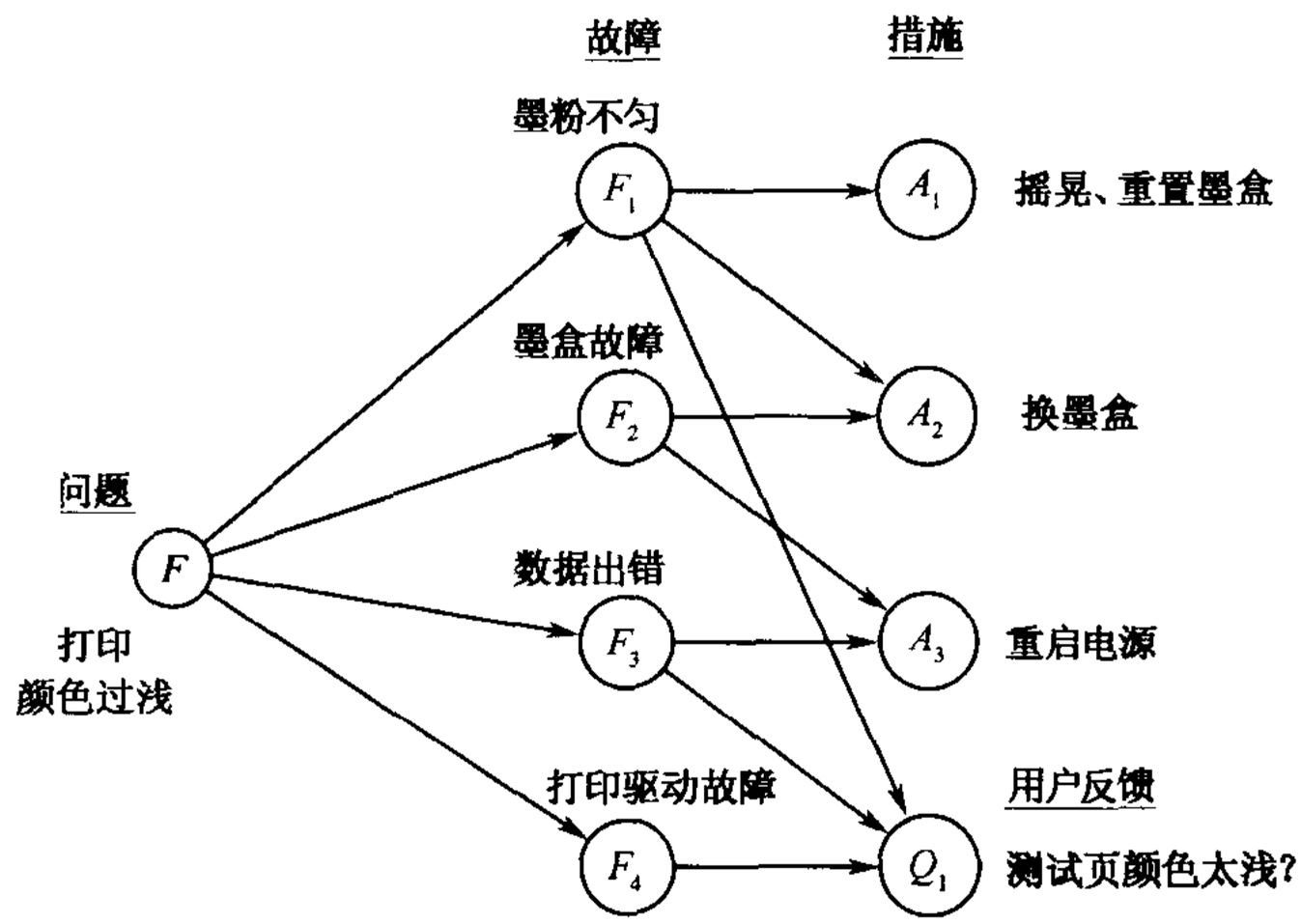


图 2.17 用于打印机故障诊断的贝叶斯网局部

target tracking)(Hautaniemi and Saarinen, 2001)、自动防御(ship self-defense)(Musman et al., 1990, 1993; Musman and Booker, 1999)、战场推理(battlefield reasoning)(Mengshoel and Wilkins, 1997; Mengshoel et al., 1998)、训练仿真(training simulation)(Grois et al., 1998)等.

战争身份识别(combat identification)是军事活动中的一个重要任务. 假设一艘航空母舰派出的作战飞机在经过战斗后要返回母舰,飞机可能因为在战斗中受损而使雷达通信和身份验证程序无法正常进行. 因此,母舰需要根据一些不确定的信息来判断飞机是敌是友,进而决定是否攻击. 传统的战争身份识别系统多是基于确定性规则的推理系统,不能很好地处理战场上的不确定性和复杂性. Laskey and Laskey (2002) 讨论了贝叶斯网在战争身份识别中的应用. 图 2.18 所示的是一个用于作战飞机身份识别的简单贝叶斯网. 其中变量"分类"是指该飞机的机型,"雷达"是该飞机的雷达信号类型,"身份"包括敌、友和中性等. 根据收集到的证据,可以计算飞机身份的后验概率分布,最终决定是否射击.

2.5.6 生态学

生态学家和野生动物学家面临的一个任务是分析人类活动对环境及濒临灭绝物种的影响. 在生态学研究中,数据的采集往往比较困难,因而需要有效地将珍贵的数据和专家的主观评估结合起来以支持有关决策,贝叶斯网为此提供了一个比较好的方法. 具体的应用实例包括:白杨林再生评估(aspen regeneration estimation)(Haas, 1991; Hass et al., 1994)、区域护林决策支持(district

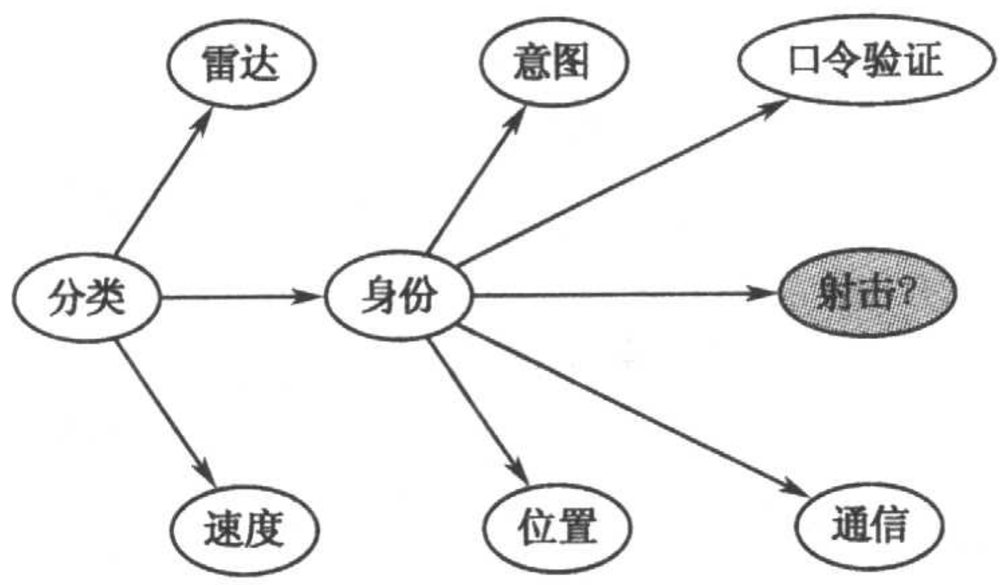


图 2.18　用于作战飞机身份识别的贝叶斯网

ranger decision making）（Haas，1992）、渔业资源管理（fishery resource management）（Lee and Rieman，1997）、人类土地利用与野生鱼类数量及栖息地的关系（Marcot et al.，2001；Rieman et al.，2001；Borsuk et al.，2002）等.

在生态学研究中，藻化现象（eutrophication）指的是由于人为因素的影响使得某个水域的藻类植物过分生长的现象. 藻化现象不仅影响美观，还会造成局部水域缺氧，使其它鱼类大量死亡，造成生态危害和经济损失. 藻化现象涉及在不同时空尺度上多个过程的相互作用，使用传统方法很难将各个过程综合在一起形成一个单一的预测模型. 贝叶斯网为解决此问题提供了可能性. Borsuk 等（2002）为 Neuse 河口藻化现象建立了一个贝叶斯网模型，如图 2.19 所示. 该网络描述了藻化现象中各变量之间的因果关系，为生态环境的监控和干涉后果的预测提供了可行的方法.

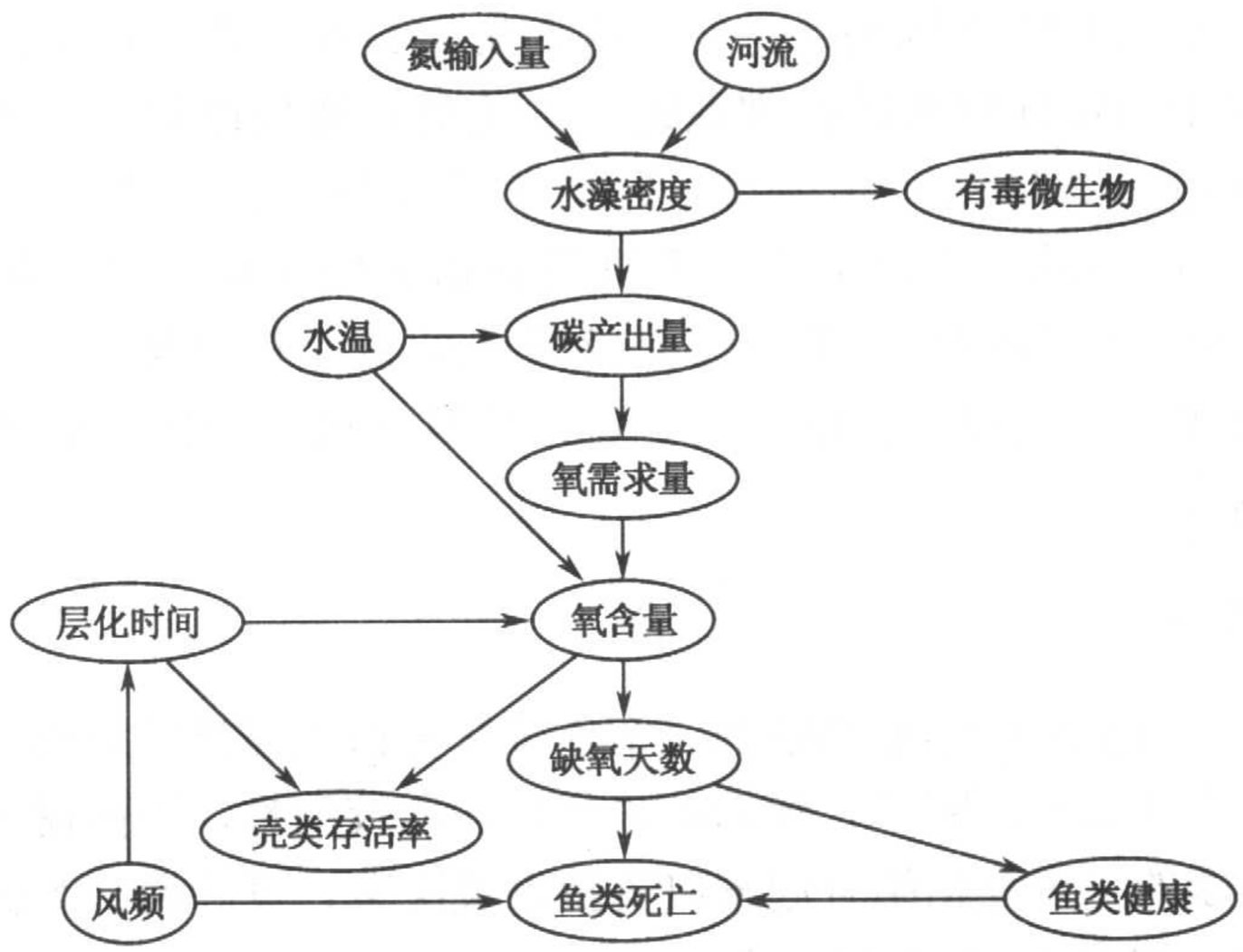


图 2.19　河口藻化现象的贝叶斯网

2.5.7 农牧业

贝叶斯网在农牧业的应用包括农作物预测、兽医诊断、农业环境分析、养殖业的动物受孕测试以及农业工程中的故障诊断等. 例如：农作物收成预测、水库资源管理决策（Said et al.，2001）、种牛血统判定（Dawid，1992；Dittmer and Jensen，1997）、母猪生育管理（Jorgensen and Toft，1999）、农业土地管理和水资源管理（Batchelor and Cain，1999；Cain et al.，2003；Bacon et al.，2002）、不使用杀虫剂的大麦麦芽生长决策支持系统（Kristensen and Rasmussen，2002）、土壤腐蚀预测（Hojsgaard et al.，2002）等.

图 2.20 是一个用于母猪生育管理的贝叶斯网的局部. 图中的 5 个节点表示了第 i 头母猪在第 t 周的情况. 其中节点 M_{ti} 表示母猪 i 的交配情况；节点 P_{ti} 表示其受孕状态；H_{ti} 和 D_{ti} 是两个观测节点，分别对应于体温测试和受孕测试. 另外，母猪的受孕率还受到其它外部过程的影响，这里用变量 Y_t 来表示. Jorgensen 和 Toft（1999）使用该贝叶斯网可以对整个养猪场的母猪生育情况进行定量分析和预测.

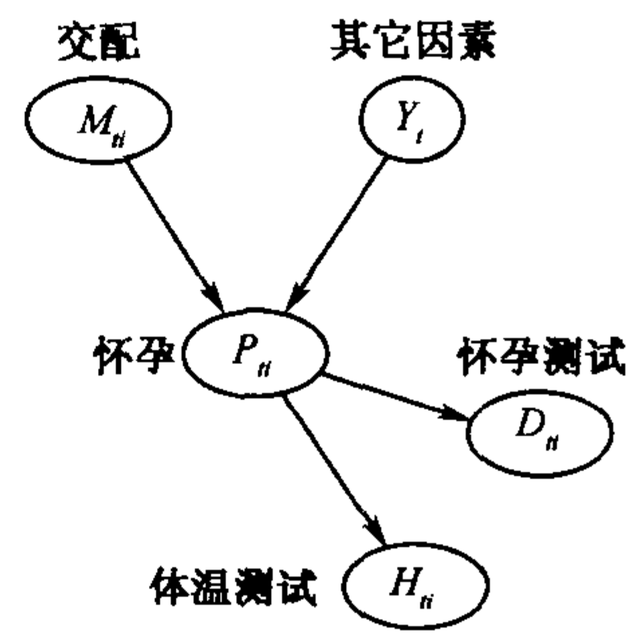


图 2.20 母猪生育管理贝叶斯网局部：第 i 头母猪第 t 周的情况

2.6 贝叶斯网对其它领域的影响

贝叶斯网的图论语言简单易懂，能够帮助人们理清问题的结构. 同时它又是严格的数学语言，可以直接用计算机来处理. 这些特点使得贝叶斯网成为许多研究领域的常用工具. 机器学习、统计学、系统工程以及模式识别中的许多模型都是贝叶斯网的特例. 本节前两小节介绍贝叶斯网在生物信息学和编码学中的使用情况，后面两小节在贝叶斯网的框架内简单介绍几个其它领域的重要模型，包括机器学习中的朴素贝叶斯模型和 TAN 模型，以及语音识别和控制领域中的隐马尔可夫模型和卡尔曼滤波器.

2.6.1 生物信息学

生物信息学是生物学与计算机科学、应用数学等相互交叉形成的一门新兴学科. 它利用计算机技术来进行生物学实验数据的获取、加工、存储、检索与分析，进而揭示数据所蕴含的生物学意义. 生物信息学是当前贝叶斯网应用最活跃的领域之一，本节简要介绍贝叶斯网在基因连锁分析中的应用（Friedman et al.，

2000；Fishelson and Geiger，2003).

细胞核中的染色体是遗传物质双螺旋分子 DNA 的载体. DNA 包含了个体的全部遗传信息，决定了个体之间的性状差异，比如身高、体重、血型、是否患有某种疾病等. 决定某一特定性状的 DNA 功能单位称为基因，它实质上是一段具有特定结构的连续的 DNA 序列，能够通过基因表达控制蛋白质的合成和细胞的分裂，对个体的性状产生影响. 我们可以抽象地把基因看成是染色体上的遗传功能单位，它们线性地排列在一条染色体上. 每个基因在染色体上的特定位置称为基因座，又称为位点（locus). 有时基因座也用来表述染色体上一个可以被识别的区域或 DNA 片段，称为 DNA 标记（marker). 同一基因的不同形式称为等位基因（allele)，通常用字母或数字表示，如 A，a，B，b，1，2，3 等. 生物个体的基因总和称为基因型（genotype)，可观察的性状总和称为表型（phenotype)，但通常所说的基因型和表型一般是针对所研究的一对或几对基因及其性状.

正常的人体细胞有两套相似的染色体（称为同源染色体)，一套来自父体，一套来自母体，共 46 条. 在配子（即精子和卵子）的产生过程中，细胞连续分裂两次，但染色体只复制一次，因此配子中的染色体数目只有正常体细胞的一半，共 23 条. 这种细胞分裂的方式叫做减数分裂. 在雌雄配子结合后，染色体的数目又恢复到正常的 46 条，保证了染色体数目的稳定性.

位于同一条染色体上的基因在遗传上一起进入同一个配子中，这种现象称为基因连锁（genetic linkage). 在同一条染色体上的基因都是相互连锁的，称为一个连锁群. 但有时也会有例外，即相互连锁的两个基因之一会与另一条同源染色体上的等位基因互换，从而使原来连锁的两个基因分别去到两个不同的配子中，这种现象称为基因互换（crossover). 基因的重组互换发生在减数分裂中，如图 2.21所示.

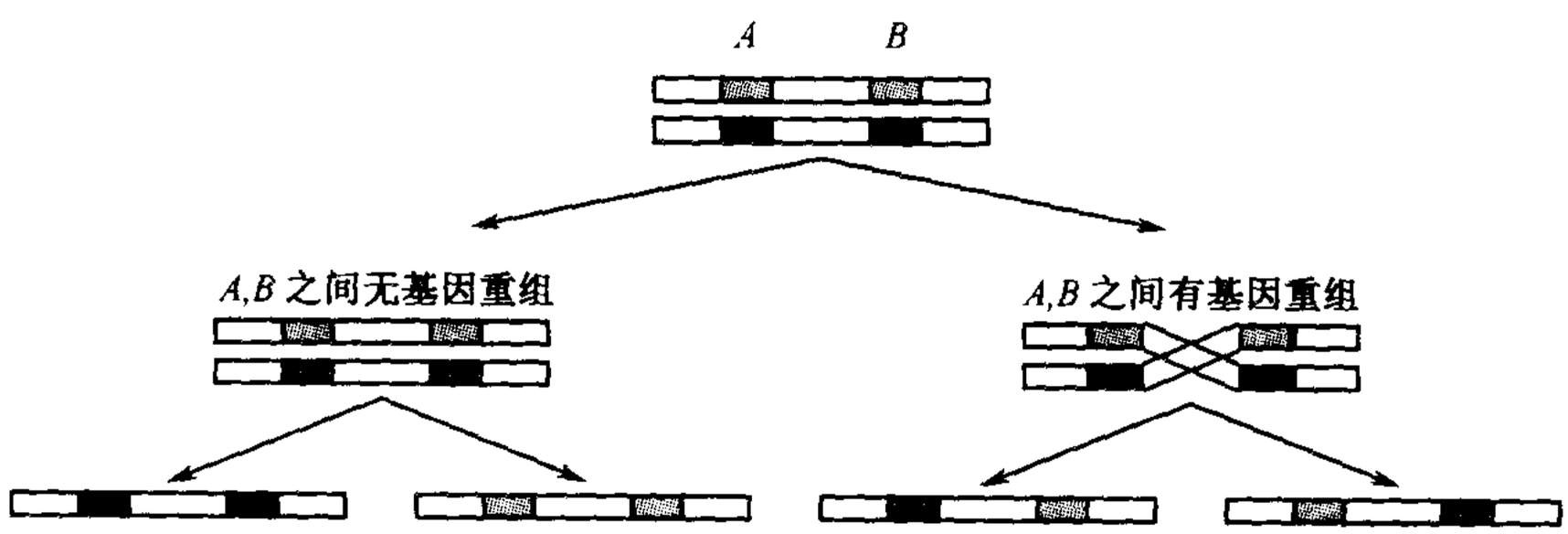


图 2.21 减数分裂时的基因重组互换示意

两个基因在染色体上的位点越接近，之间发生交换的概率就越小，反映出它们之间的连锁紧密. 反之，两个基因相距越远，之间发生交换的概率就越大，反

映出它们之间的连锁较弱. **基因连锁分析**（genetic linkage analysis）就是根据这一原理来推测决定某种表型的基因在染色体上的位置. 基因连锁分析的主要目标是得到一张基因连锁图（genetic linkage map），图中基因与基因之间或基因与遗传标记之间的距离以厘摩（cM，centiMorgan）为单位①. 染色体上两个位点之间的遗传距离为1个厘摩指的是它们在减数分裂中发生重组的概率为1%，即每100次分裂中发生1次重组②.

在基因连锁分析中，无法直接观察到染色体上的基因型是什么，而只能观察到个体的表型，如血型，肤色，是否有某种疾病等. 在对某个致病基因进行研究时，通常先收集病人整个家系的相关信息，得到表型数据 $\mathscr{D}$，然后画出一个如图2.22所示的家系树（pedigree）. 家系树描述了家系中个体之间基因型和表型的遗传关系，定义了所有个体的基因型和表型的一个联合概率分布. 假设个体 i 的基因型和表型分别为 $G[i]$ 和 $P[i]$，则根据遗传学知识，可以得到以下独立关系：①给定父母的基因型 $G(\pi(i))$，个体 i 的基因型 $G[i]$ 独立于任何其它祖先 j 的基因型 $G(j)$；②给定基因型 $G[i]$，个体 i 的表型 $P[i]$ 独立于所有其它基因型和表型变量. 另外，我们还需引入一个“选择变量”来表示家系树中的遗传模式，以描述减数分裂过程中个体 i 的基因重组和交换的不确定性.

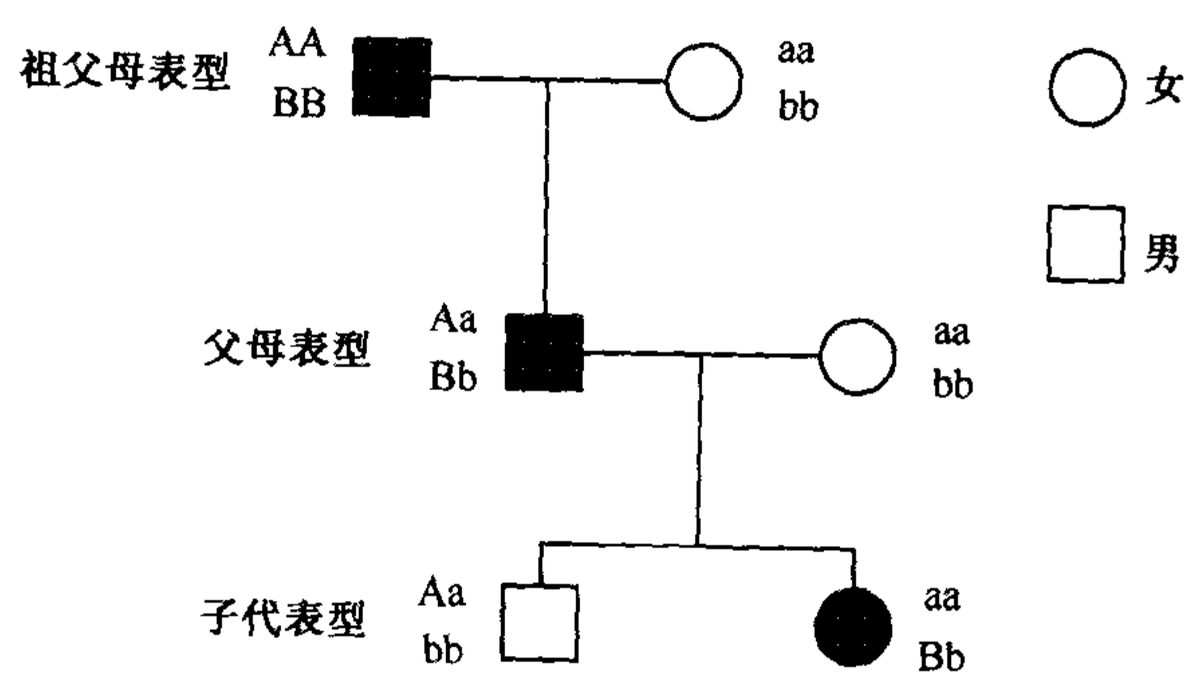


图2.22 一个包含6个个体、2个位点的三代家系树
（其中方块为男性，圆圈为女性，黑色表示患病个体，白色表示正常个体）

利用上述假设，可以把一个家系树表达为一个参数化的贝叶斯网，其中的参数是待确定位置的致病基因与某两个已知位点之间的基因重组率. 我们假设未知致病基因位于某两个已知的遗传标记或基因之间，画出相应的家系树，并将其转化为一个贝叶斯网，然后针对这个贝叶斯网计算表型数据的似然度. 可以对未知

① 以此纪念现代遗传学的先驱摩尔根（Thomas Hunt Morgan，1866～1945）.

② 人类基因组的平均遗传长度为3300cM，而DNA的平均长度为30亿个碱基对，因此1cM相当于1000000个碱基对的物理长度.

致病基因的每个可能位置和每个可能的基因重组率都进行上述计算，最后取使表型数据的似然率达到最大的那个位置和基因重组率.

图 2.23 所示为一个包括 3 个个体（父、母、子）和 3 个基因位点的家系树（Friedman et al.，2000），其中变量 A, B, C 为基因型，变量 F, G, H 为表型，S 为选择变量. $A[i,f]$ 表示来自个体 i 的父亲的基因，$A[i,m]$ 表示来自个体 i 的母亲的基因. 网络中选择变量 S 的条件分布表包括待定的参数——基因重组率，其它各基因型和表型变量的条件分布表可以利用遗传学的知识来确定，详见文献（Friedman et al.，2000）. 图 2.23 暗含了一个关于基因 A, B, C 在染色体上的物理顺序的假设，这反映在选择变量 $S_A[i,p]$ 和 $S_C[i,p]$ 之间没有直接的边相连，即给定 $S_B[i,p]$ 的取值，$S_A[i,p]$ 和 $S_C[i,p]$ 之间相互独立. 可以想象 B 是待定位置的致病基因，A 和 C 是已知位置的基因或标记. 对这个贝叶斯网，我们可以尝试 A 和 B 之间以及 B 和 C 之间的基因重组率的多个取值组合，对每个组合计算数据的似然度 $P(D)$，最后取使似然度达到最大的那个组合. 在这里，基因连锁分析问题被转化为贝叶斯网中的推理问题，常用的算法是后面将会介绍到的变量消元法. 当问题包含多个个体和多个基因时，对应网络的规模可能很大，达到上千个节点.

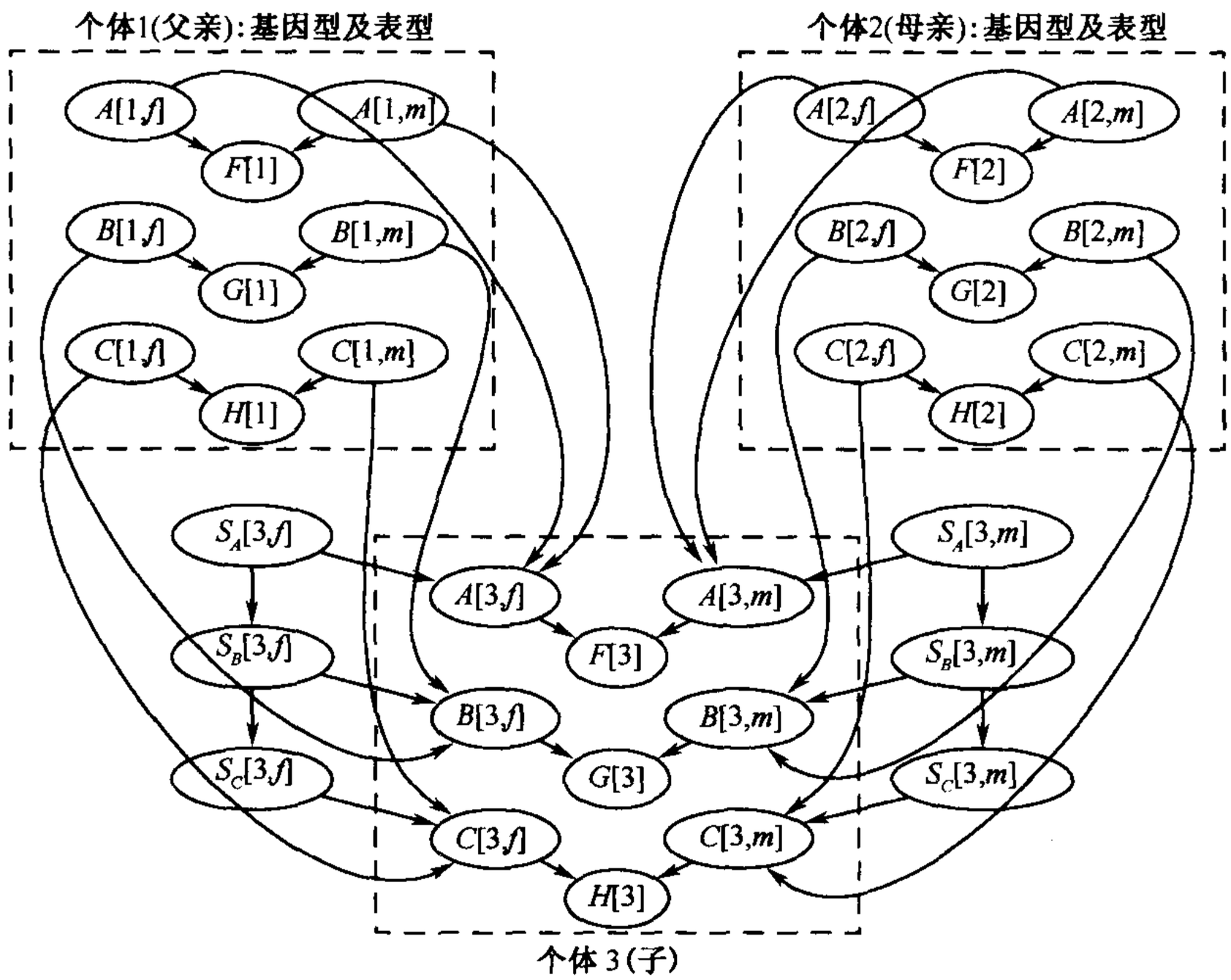


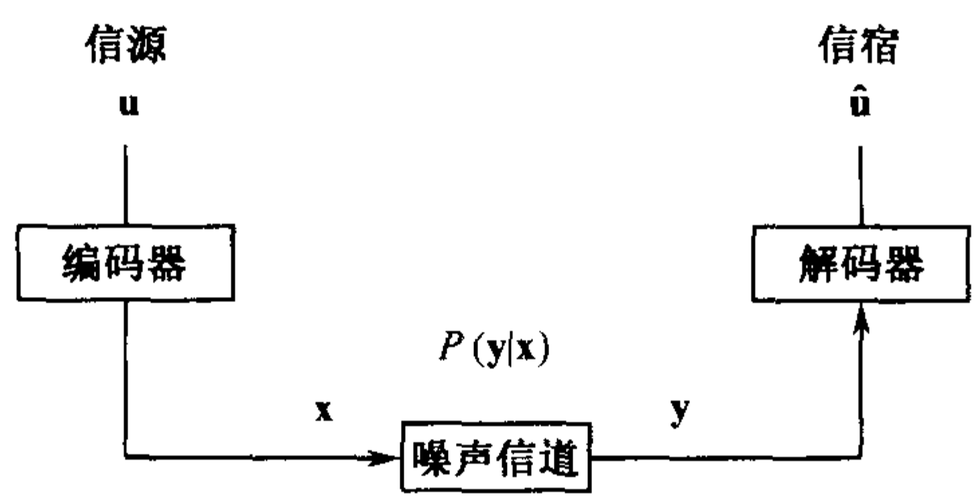
图 2.23　包含 3 个个体和 3 个基因位点的家系树贝叶斯网

A, B, C:基因型;F, G, H:表型;S:选择变量

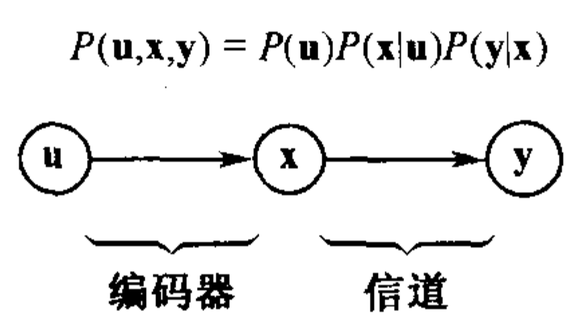
除基因连锁分析以外，贝叶斯网在生物信息学中的应用还有很多，例如：基因表达数据的分析（Friedman et al.，2000；Spirtes et al.，2000）、法庭调查中的 DNA 身份验证（forensic identification）（Dawid et al.，2002）、基因变异分析（Dawid，2003）、基因微阵列质量控制（microarray quality control）（Hautaniemi et al.，2003）、遗传操纵子预测（operon prediction）（Bockhorst et al.，2003）以及进化树分析（phylogenetic tree analysis）（Strimmer and Moulton，2000；Friedman et al.，2002）等.

2.6.2　编码学

数字通信的首要问题是如何通过**信道编码**（channel coding）来实现噪声信道上的无错传输. 这里编码指的是按照一定的规则为信源传输序列 **u** 加入冗余信息，以得到编码后的序列 **x**，然后把 **x** 通过噪声信道进行传输. 接收端收到一个因干扰而包含错误的接收序列 **y**，对 **y** 进行解码，以得到尽可能接近原传输序列 **u** 的信宿 $\hat{\mathbf{u}}$.



(a) 信道编码模型



(b) 信道编码贝叶斯网

图 2.24　信道编码模型及其贝叶斯网表示

在如图 2.24(a)所示的通信系统模型中，对 **y** 进行解码之前的过程对应了如下的一个联合概率分布的分解：

$$P(\mathbf{u},\mathbf{x},\mathbf{y}) = P(\mathbf{u})P(\mathbf{x} \mid \mathbf{u})P(\mathbf{y} \mid \mathbf{x}). \tag{2.8}$$

这样的概率模型可以用一个如图 2.24(b)所示的贝叶斯网来描述. 这样，解码问题就转化为贝叶斯网上的推理问题：已知 **y** 时，计算使后验概率 $P(\mathbf{u} \mid \mathbf{y})$ 达到最

大的传输序列为

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u}} P(\mathbf{u} \mid \mathbf{y}). \tag{2.9}$$

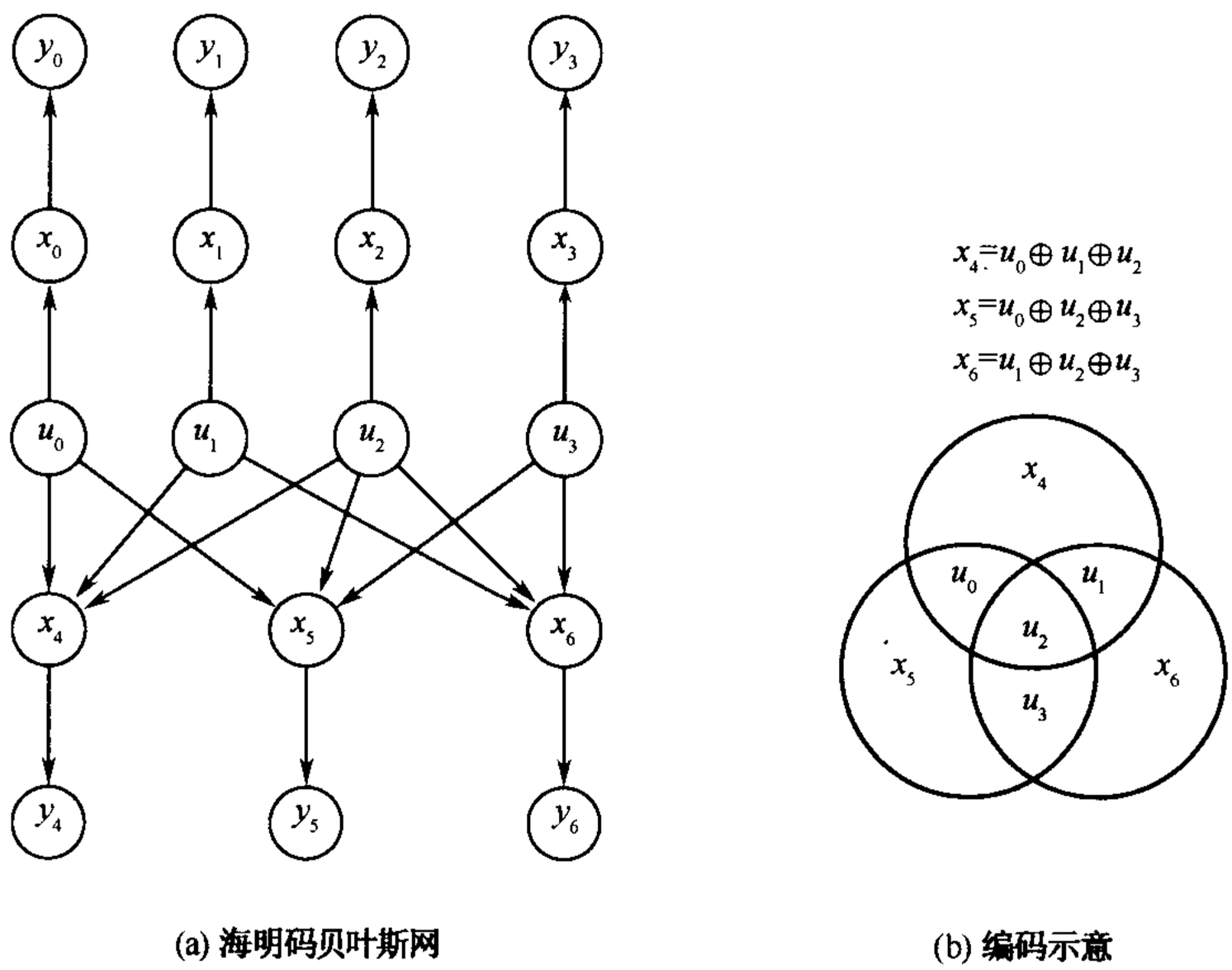
下面讨论两个典型的信道编码的例子.

1. 海明码

海明码（Hamming code）是一类简单的纠错码. 它往传输序列 $\mathbf{u}$ 加入多个校验位来生成编码序列 $\mathbf{x}$，其中每个校验位都依赖于 $\mathbf{u}$ 的一个子集. 用 K 和 N 分别记 $\mathbf{u}$ 和 $\mathbf{x}$ 的长度，一个 (N,K) 海明码对长度为 K 的二进制传输序列编码，生成一个长度为 N 的编码序列，其编码效率为 K/N. N 和 K 必须满足 $N = 2^m - 1$ 和 $K = 2^m - m - 1$，其中 m 是某个不小于 2 的整数. 下面看一个具体的例子. 图 2.25(a)中的贝叶斯网络表示了一个利用（7，4）海明码进行信道编码的通信模型. 其中 $\mathbf{u} = u_0u_1u_2u_3$ 是一个长度为 4 的二进制传输序列，$\mathbf{x} = x_0x_1x_2x_3x_4x_5x_6$ 是通过海明码编码得到的码字（codeword），$\mathbf{y} = y_0y_1y_2y_3y_4y_5y_6$ 是在噪声信道的接收端得到的接收序列. 假设传输序列 $\mathbf{u}$ 的各位是随机且独立产生的，则有 $P(u_k = 0) = P(u_k = 1) = 0.5$，$k = 0,1,2,3$. 码字 $\mathbf{x}$ 的前 4 位是对 $\mathbf{u}$ 的简单复制，所以 $P(x_k \mid u_k) = \delta(x_k, u_k)$，$k = 0,1,2,3$，这里 δ 函数定义如下：

$$\delta(a,b) = \begin{cases} 1, & \text{如果 } a = b; \\ 0, & \text{其它}. \end{cases}$$

其它 3 个校验位的产生规则如图 2.25(b)所示，即



(a) 海明码贝叶斯网　　(b) 编码示意

图 2.25　（7，4）海明码及其对应的贝叶斯网

$$P(x_4 \mid u_0, u_1, u_2) = \delta(x_4, u_0 \oplus u_1 \oplus u_2),$$

$$P(x_5 \mid u_0, u_2, u_3) = \delta(x_5, u_0 \oplus u_2 \oplus u_3),$$

$$P(x_6 \mid u_1, u_2, u_3) = \delta(x_6, u_1 \oplus u_2 \oplus u_3).$$

最后，$P(y_i \mid x_i)$ 取决于所采用的传输信道，常见的有**二值对称信道**（binary symmetric channel）和**加性高斯白噪声**（additive white Gaussian noise，AWGN）信道，详见文献（Frey，1998；Mackay，2004）.

2. Turbo 码

Shannon 的信道编码定理指出：只要传输率低于信道容量，就总存在使得出错率任意小的编码方法（Shannon，1948）. 也就是说，只要传输信噪比不低于 Shannon 极限，就总存在能使出错率任意小的编码方法. Shannon 的理论证明了这样的码的存在，却没有指出实际中应该怎样实现它. 自 Shannon 的理论发表以来，人们为设计有效的纠错码做了大量的努力. 1993 年，C. Berrou 等提出了 Turbo 码（Berrou et al.，1993），获得了接近 Shannon 理论极限的性能，成为编码领域一个激动人心的重要发现.

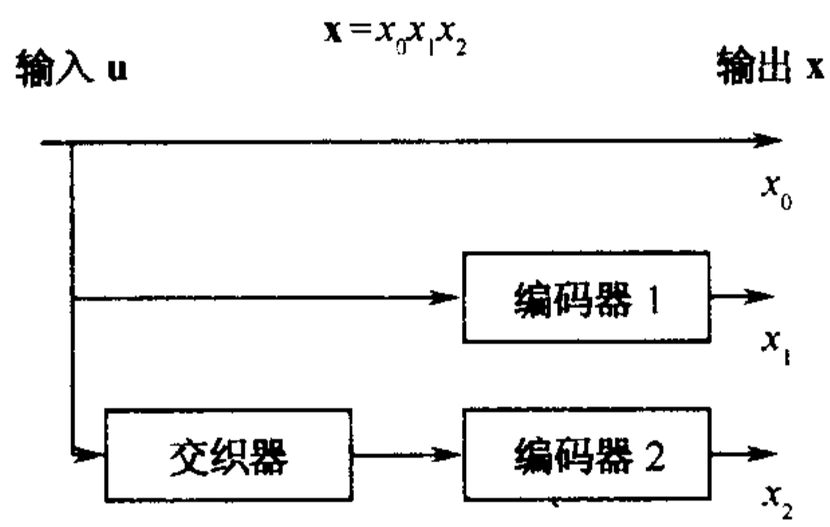
Turbo 码又称为**并行级联卷积码**（parallel concatenated convolutional code），它并行地使用两个一样的卷积码进行编码，但在其中一个编码器之前加了一个**随机交织器**（random interleaver），以在编码前对传输序列随机排序. Turbo 码的编码部分如图 2.26(a)所示. 图 2.26(b)中的贝叶斯网表示了一个利用 Turbo 码进行信道编码的通信模型. 其中传输序列 $\mathbf{u}$ 的长度为 6，编码后的码字 $\mathbf{x}$ 的长度为 12. 注意二值变量 s_i^j 是在 Turbo 码中起辅助作用的中间变量. 在引入中间变量 $\mathbf{s}$ 后，信道编码的概率模型[①]为

$$P(\mathbf{u},\mathbf{s},\mathbf{x},\mathbf{y}) = P(\mathbf{u})P(\mathbf{s} \mid \mathbf{u})P(\mathbf{x} \mid \mathbf{u},\mathbf{s})P(\mathbf{y} \mid \mathbf{x}). \tag{2.10}$$

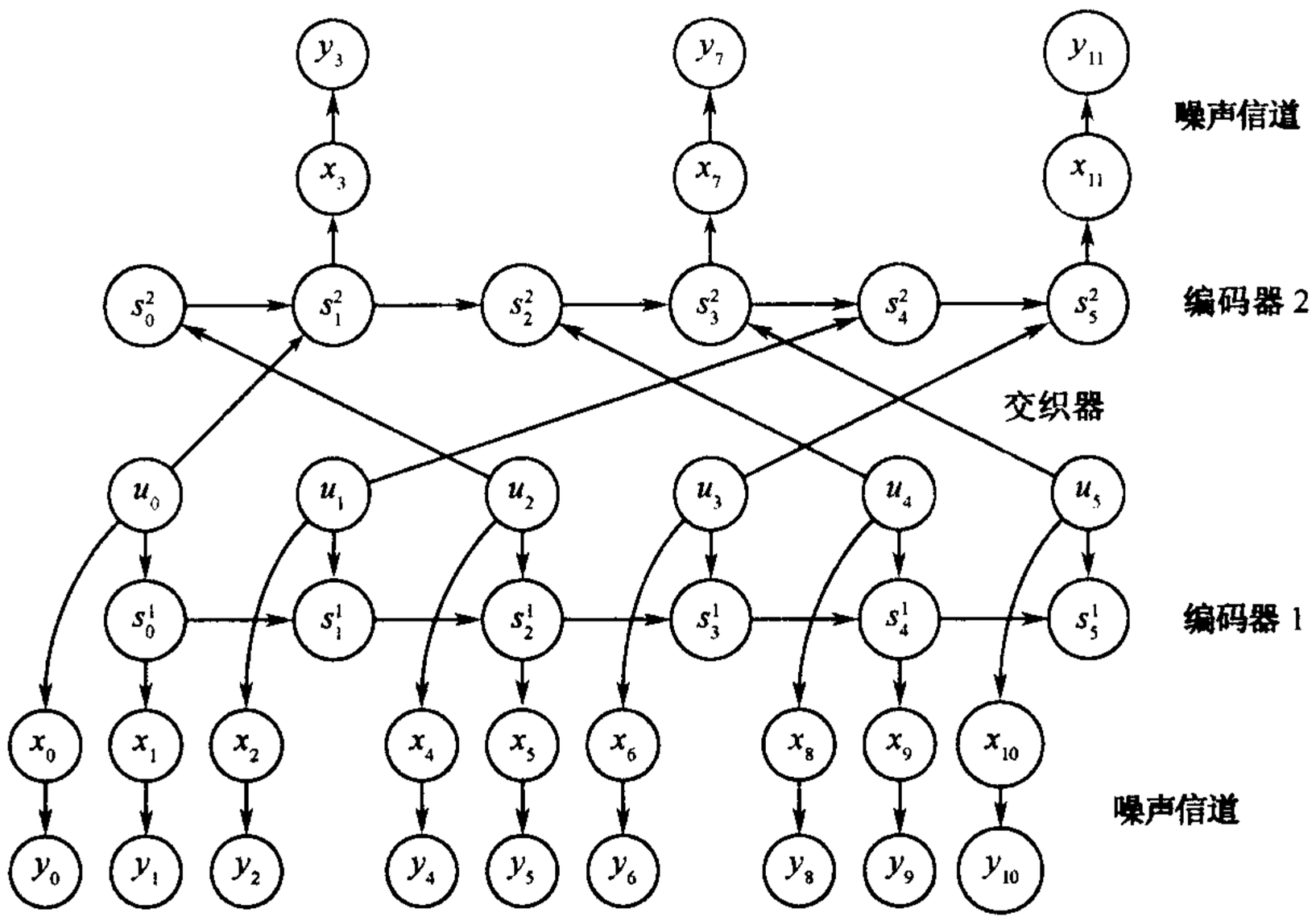
它可以用一个如图 2.26(c)所示的贝叶斯网来描述.

Turbo 码最突出的特点是它的迭代解码算法：两个卷积解码器相互传递“消息”，交换各自的中间解码结果，系统经过数次迭代后（通常小于 20 次），得到最终的解码结果. McEliece 等（1998）指出，Turbo 码的迭代解码算法实际上就是 Pearl 的信度传播（belief propagation）贝叶斯网推理算法的一个特例. 这个发现揭示了纠错码和贝叶斯网之间的密切联系，为纠错码的研究开辟了新的方向. 关于 Turbo 码和贝叶斯网之间关系的更多讨论参见有关文献（Mackay，

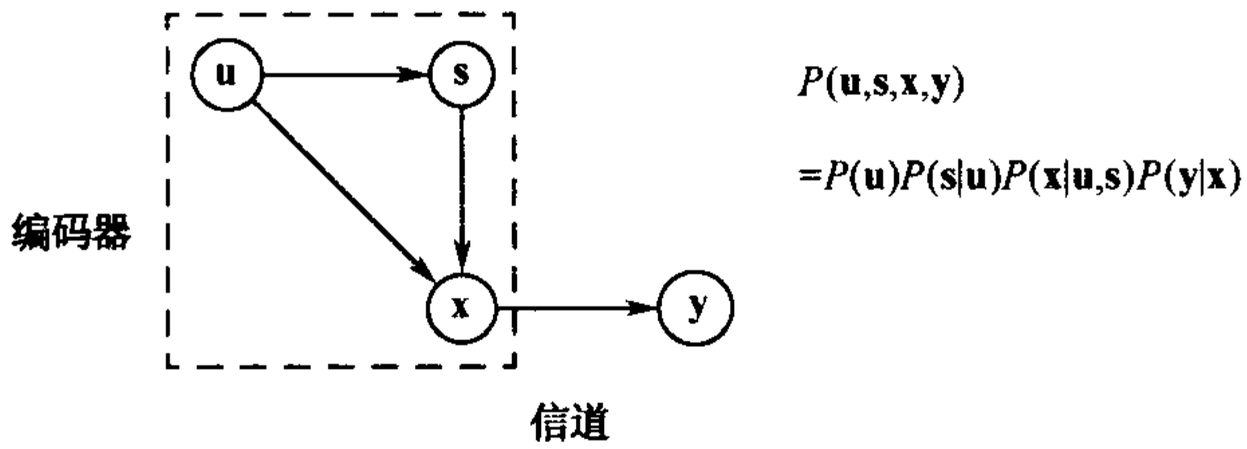
① 网络中各个参数的设定见文献（Frey，1998）.



(a) Turbo 码编码器示意



(b) Turbo 码贝叶斯网: K=6,N=12



(c) 引入中间变量 s 的信道编码贝叶斯网

图 2.26　Turbo 码及其对应的贝叶斯网

2004; Frey and Mackay, 1998; Kschischang et al., 2001; Kschischang and Frey, 1998; Deng, 1999).

2.6.3 机器学习

机器学习是人工智能的一个子领域，它研究的是怎样让计算机模拟实现人的学习行为，从经验数据出发，通过学习不断提高自己完成某项任务的能力. **分类**(classification) 和**聚类** (clustering) 是机器学习中的两个主要问题. 分类是从一系列给定类别信息的数据出发，为下一个未知类别的数据归类. 聚类是从一系列未知类别信息的数据出发，分析它们可以聚成几个类，以及哪些数据属于哪一类. 本小节简单介绍两个常见的分类模型，它们都是贝叶斯网的特例，即朴素贝叶斯模型和 TAN 模型.

1. 朴素贝叶斯模型

朴素贝叶斯模型 (naïve Bayes model)，又称**朴素贝叶斯分类器** (naïve Bayes classifier)，是一个包含一个根节点、多个叶节点的树状贝叶斯网，如图 2.27 所示. 其中叶节点 $A_1,\cdots,A_n$ 是属性变量，描述待分类对象的属性，根节点 C 是类别变量，描述对象的类别. 用朴素贝叶斯模型进行分类就是给定一个数据点，即各属性变量的取值 $A_1=a_1,\cdots,A_n=a_n$，计算后验分布 $P(C\mid A_1=a_1,\cdots,A_n=a_n)$，然后选择概率最大的那个 C 值作为这个数据点所属的类别. 在医疗诊断中，C 代表一系列疾病，$A_1,\cdots,A_n$ 代表这些疾病可能导致的症状，诊断就是根据症状来确定疾病.

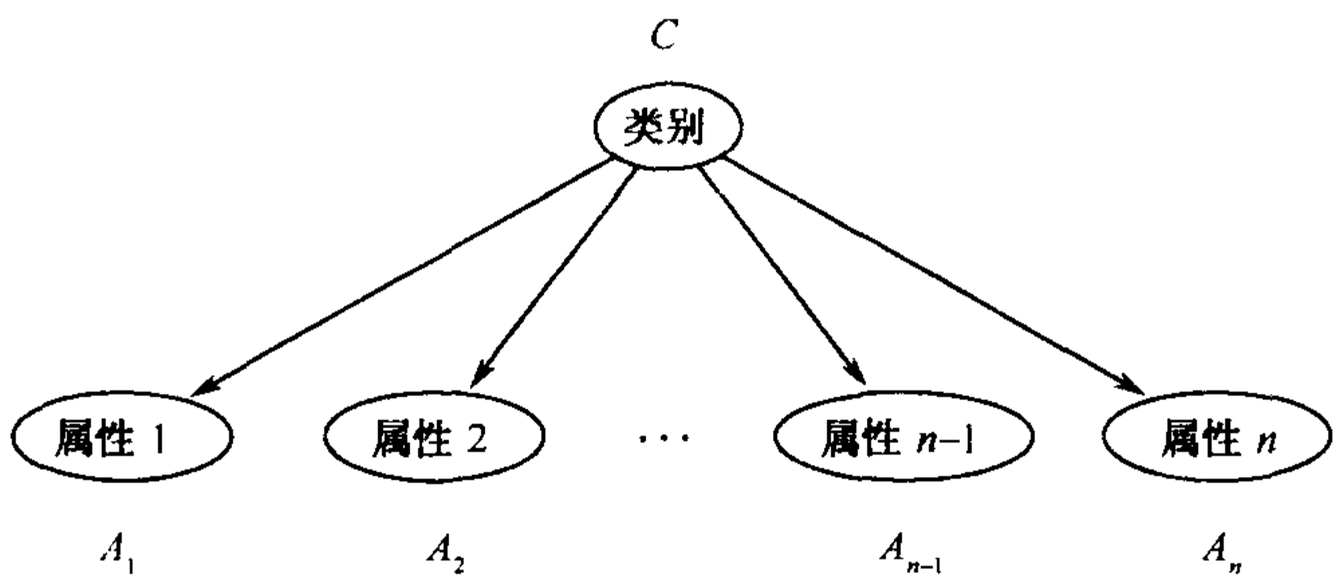


图 2.27 朴素贝叶斯分类器

朴素贝叶斯模型包含了一个所谓的**局部独立** (local independence) 假设，即给定类别变量 C，各属性变量 A_i 相互条件独立. 这意味着，联合概率分布满足下式：

$$P(C,A_1,\cdots,A_n) = P(C)\prod_{i=1}^{n} P(A_i \mid C). \tag{2.11}$$

朴素贝叶斯模型最早由 Warner 等（1961）提出，用于先天性心脏病的诊断，其主要优点是结构简单，计算复杂度低. 尽管它所采用的局部独立假设显然过于理想化，但实验数据显示，其性能和其它常见的分类器不相上下（de Dembal et al.，1972；Duda and Hart，1973；Adams et al.，1986；Langley et al.，1992).

2. TAN 模型

朴素贝叶斯模型中的局部独立假设在实际中往往不成立. 为使模型更符合实际，可以在其中的各叶节点之间增加一些必要的边，以表示各属性变量之间的依赖关系，如果规定属性变量之间的关系成树状结构，那么模型就称为**加树朴素贝叶斯模型**（tree augmented naïve Bayes model)，简称 **TAN 模型**. 图 2.28 给出了这样的一个例子. 在 TAN 模型中，联合概率分布为

$$P(C,A_1,\cdots,A_n) = P(C)\prod_{i=1}^{n} P(A_i \mid \pi(A_i)). \tag{2.12}$$

注意这里的属性变量 A_i 的父节点 $\pi(A_i)$ 不仅包括类别变量 C，也可能包括其它属性变量. 相比之下，在朴素贝叶斯模型中 $\pi(A_i)$ 只包括 C. 这说明，TAN 模型不再要求局部独立假设成立. 有研究表明，TAN 模型的分类效果往往比朴素贝叶斯要好（Friedman et al.，1997；Cheng and Greiner，1999).

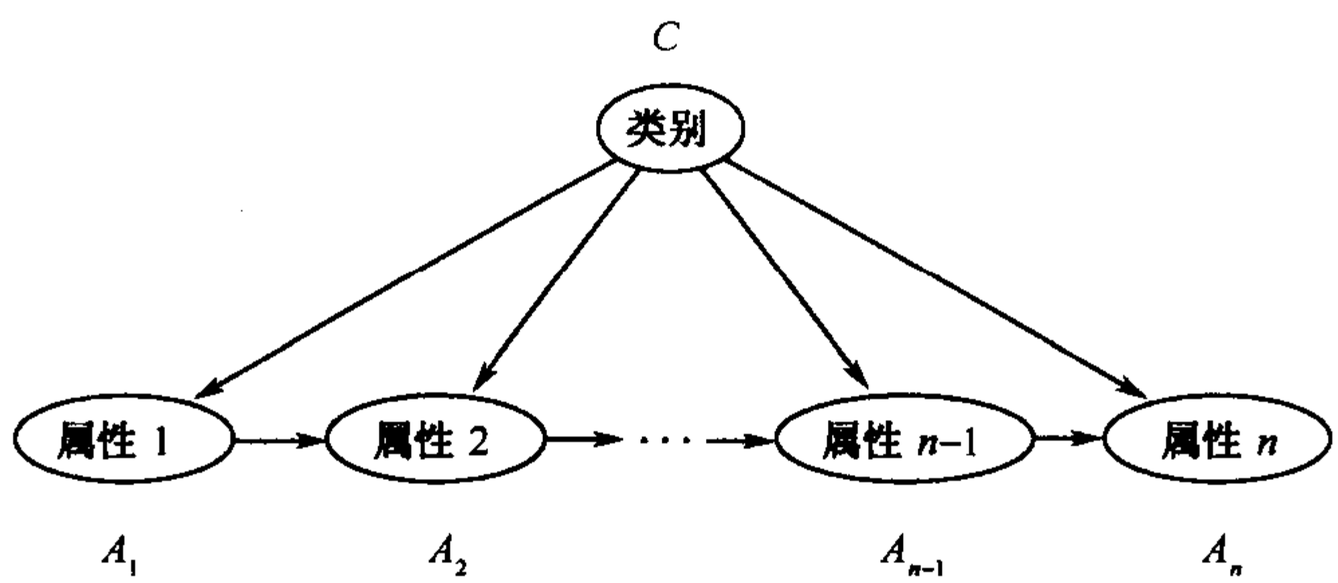


图 2.28 TAN 模型

2.6.4 时序数据和动态模型

实际中的许多随机现象都涉及一些随时间变化的随机变量，如植物的生长、语音的产生以及连续变化的视觉图像等. 为了能够对此类动态时变随机过程进行表达和推理，人们引入了**动态贝叶斯网**（dynamic Bayesian networks，DBN)

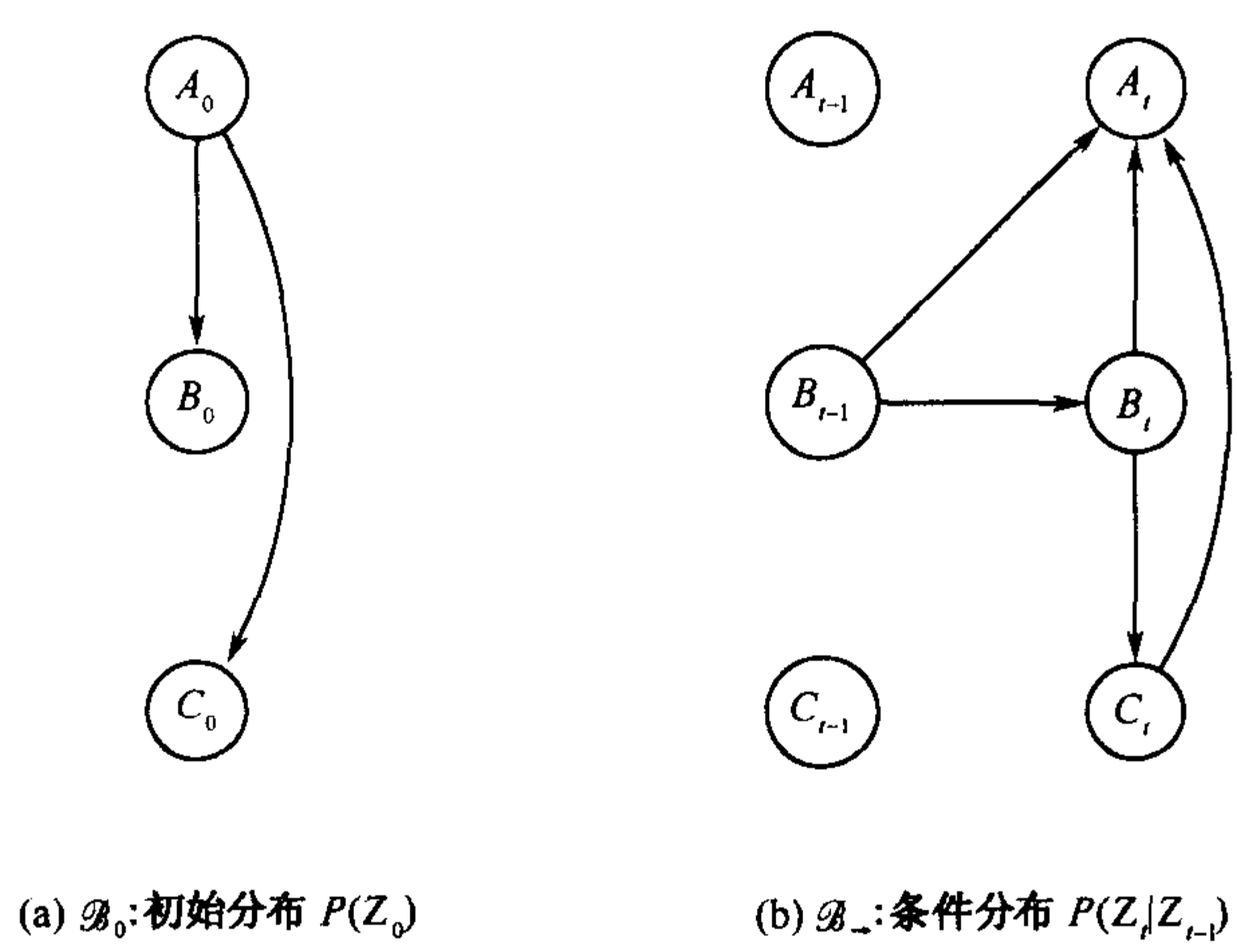
(Dean and Kanazawa, 1990; Russell and Norvig, 2003) 的概念，它又称为**时变贝叶斯网** (temporal Bayesian networks).

一个动态贝叶斯网可以定义为 $(\mathscr{B}_0, \mathscr{B}_{\rightarrow})$，其中 $\mathscr{B}_0$ 是一个标准贝叶斯网，定义了初始时刻的概率分布 $P(Z_0)$，$\mathscr{B}_{\rightarrow}$ 是一个包含两个时间片的贝叶斯网，定义了两个相邻时间片的各变量之间的条件分布，即

$$P(Z_t \mid Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^i \mid \pi(Z_t^i)). \tag{2.13}$$

其中 Z_t^i 是位于时间 t 时的节点 i，$\pi(Z_t^i)$ 是 Z_t^i 的父节点. $\mathscr{B}_{\rightarrow}$ 中前一个时间片中的节点可以不给出参数，第二个时间片中的每个节点都有一个条件概率分布 $P(Z_t^i \mid \pi(Z_t^i)), t > 0$. 节点 Z_t^i 的父节点 $\pi(Z_t^i)$ 可以在同一时间片内，也可以在前一时间片内. 位于同一时间片内的边可以理解为瞬时作用，而跨越时间片的边可以理解为时变作用，反映了时间的流逝.

例 2.8 图 2.29 中所示的是一个动态贝叶斯网，其中图(a)是 $\mathscr{B}_0$，定义了初始时刻的分布 $P(Z_0)$，图(b)是 $\mathscr{B}_{\rightarrow}$，定义了两相邻时间片之间的条件概率 $P(Z_t \mid Z_{t-1})$，每个时间片包含 3 个节点. □



(a) $\mathscr{B}_0$:初始分布 $P(Z_0)$　　(b) $\mathscr{B}_{\rightarrow}$:条件分布 $P(Z_t|Z_{t-1})$

图 2.29 动态贝叶斯网示例：每个时间片包含 3 个节点

动态贝叶斯网包含了两个假设：①一阶马尔可夫假设，即各节点之间的边或者位于同一时间片内，或者位于相邻时间片之间，不能跨越时间片；②时齐性或齐次性，即 $\mathscr{B}_{\rightarrow}$ 中的参数不随时间变化. 根据初始分布和相邻时间片之间的条件分布，可以将动态贝叶斯网展开到第 T 个时间片，结果得到一个跨越多个时间片的联合概率分布，即

$$P(Z_{0:T}) = \prod_{t=0}^{T} \prod_{i=1}^{N} P(Z_t^i \mid \pi(Z_t^i)). \tag{2.14}$$

下面介绍两个动态贝叶斯网的特例，即隐马尔可夫模型和卡尔曼滤波器.

1. 隐马尔可夫模型

隐马尔可夫模型（HMM）是一个简单的时间序列模型. 它的每个时间片由一个隐状态变量 X_t 和一个观测变量 Y_t 组成，它们都是离散变量. 隐马尔可夫模型的名字正好反映了它的两个性质：首先，它假设时刻 t 时的观测值是由某个其状态隐藏的过程产生的；其次，这个隐藏的状态变量满足所谓的马尔可夫假设，即给定 X_{t-1} 的值，X_t 与所有 $t-1$ 时刻之前的状态无关. 图 2.30 所示的是一个隐马尔可夫模型的前 4 个时间片.

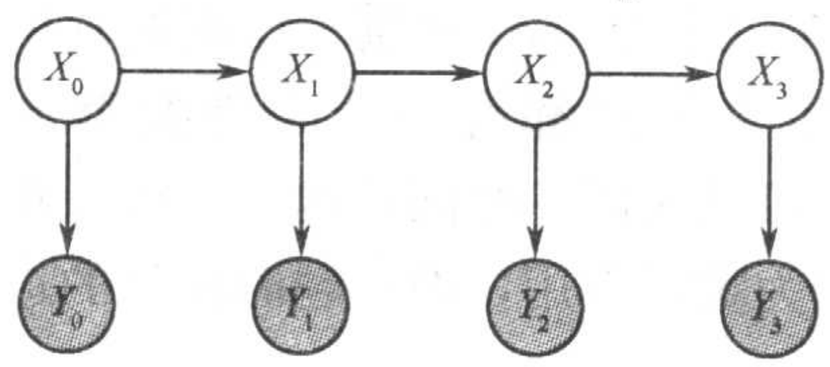


图 2.30 HMM 示例：前四个时间片

HMM 被广泛应用于语音识别（Rabiner，1989）和生物序列分析（Birney，2001）领域. 显然，HMM 是一个简单的动态贝叶斯网，其中

$$\mathscr{B}_0: P(Z_0) = P(X_0, Y_0) = P(X_0)P(Y_0 \mid X_0), \tag{2.15}$$

$$\mathscr{B}_{\rightarrow}: P(Z_t \mid Z_{t-1}) = P(X_t, Y_t \mid X_{t-1}, Y_{t-1}) = P(X_t \mid X_{t-1})P(Y_t \mid X_t). \tag{2.16}$$

两者之间关系的详细介绍参见有关文献（Ghahramani，2001；Murphy，2002；Jordan，2004；Russell and Norvig，2003）.

2. 卡尔曼滤波器

卡尔曼滤波器（KFM），又称**线性动态系统**（linear dynamic system），由 Kalman（1960）提出，并被用于解决控制和通信领域中的线性滤波和预测问题. KFM 涉及一个随时间变化的连续变量 x，x 随时间的演变以及对 x 的观测结果 y 都符合线性高斯分布，即 x_{t+1} 是 x_t 的线性函数再加一个高斯过程噪声，y_t 是 x_t 的线性函数再加一个高斯测量噪声，即

$$x_{t+1} = Ax_t + w_t, \tag{2.17}$$

$$y_t = Bx_t + v_t, \tag{2.18}$$

其中 $w \sim N(0, Q)$ 是高斯过程噪声，$v \sim N(0, R)$ 是高斯测量噪声.

KFM 的结构和 HMM 类似，唯一不同的是 HMM 中的状态变量是离散的，而 KFM 中的状态变量和观测变量都是连续的，且服从线性高斯分布. KFM 也是动态贝叶斯网的一个特例，是一个简单的连续型动态贝叶斯网. 它在控制和信号处理领域应用很广，包括飞机导弹的雷达跟踪控制、视觉或声纳反射目标的跟

踪、信号滤波等（Kalman，1960；Minka，1999；Roweis and Ghahramani，1999；Murphy，2002；Russell and Norvig，2003）.

2.7 文献介绍

贝叶斯网的概念由 Pearl（1986）首先提出. Pearl 的专著《Probabilistic Reasoning in Intelligent Systems：Networks of Plausible Inference》于 1988 年面世，随后又有几本关于贝叶斯网的著作相继出版（Shafer，1996；Lauritzen，1996；Jordan，1998；Cowell et al，1999；Jensen，2001；Neapolitan，2003）. 与上述著作相比，本书系统、简洁地介绍贝叶斯网常用的概念和结果，更适合作为入门教材使用. 读者若想进一步了解某些专题，可以参考上述书籍.

从统计学角度来看，贝叶斯网是**图模型**（graphical models）的一种. 图模型在 20 世纪初期已出现在遗传学（Wright，1921）、物理学（Gibbs，1902）以及概率论（Markov，1912）的文献中. 近期也有综述文章（Wermuth，1998；Wermuth and Cox，2001 和专著（Edwards，2000；Lauritzen，1996）问世.

决策网（decision networks）是另一个与贝叶斯网紧密相关的模型. 贝叶斯网使得人们可以把概率论应用于复杂领域，而决策网使得人们可以把决策论（decision theory）应用于复杂领域. 决策网中有 3 类不同的节点：表示决策变量的**决策节点**（decision node）、表示效用函数的**效用节点**（utility node）以及表示随机变量的**机会节点**（chance node）. 相对而言，贝叶斯网中只有表示随机变量的机会节点. 决策网有时也被称为 influence diagram，最早由 Howard 和 Matheson（1984）提出，被引用最多的文章是《Evaluating influence diagram》（Shachter，1986）. Jensen（2001）和 Cowell 等（1999）对于决策网都有专门论述. Zhang 等（1994）和 Zhang（1998b）指出 influence diagram 中的某些限制是不必要的，从而对它进行了推广，并明确了决策网中的决策优化问题与贝叶斯网中的推理问题之间的关系.

第 3 章　图分隔与变量独立

贝叶斯网是概率论与图论相结合的产物. 在一个贝叶斯网中，一方面可以从概率论的角度谈论变量之间的依赖与独立，另一方面也可以从图论的角度谈论节点之间的连通与分隔. 本章揭示贝叶斯网的概率论侧面与它的图论侧面之间的深刻联系.

3.1　直观分析

在这一节里，我们首先通过直观分析引出一个在贝叶斯网中判别变量间条件独立关系的图论准则，即有向分隔. 下一节将严格证明这一准则的正确性.

在贝叶斯网中，两个变量 X 和 Y 如果直接相连，则表示它们之间有直接依赖关系，对 X 的了解会影响关于 Y 的信度，反之亦然. 在图 3.1 所示的 Alarm 网络中，盗窃 (B) 与警铃响 (A) 之间有一条边 "$B \to A$"，表示 A 对 B 有直接依赖关系：若知道"发生了盗窃"，则对"警铃响了"的信度会增加；同样，若知道"警铃响了"，则对"发生了盗窃"的信度也会增加. 在这种意义下，我们称信息能够在两个直接相连的节点之间传递.

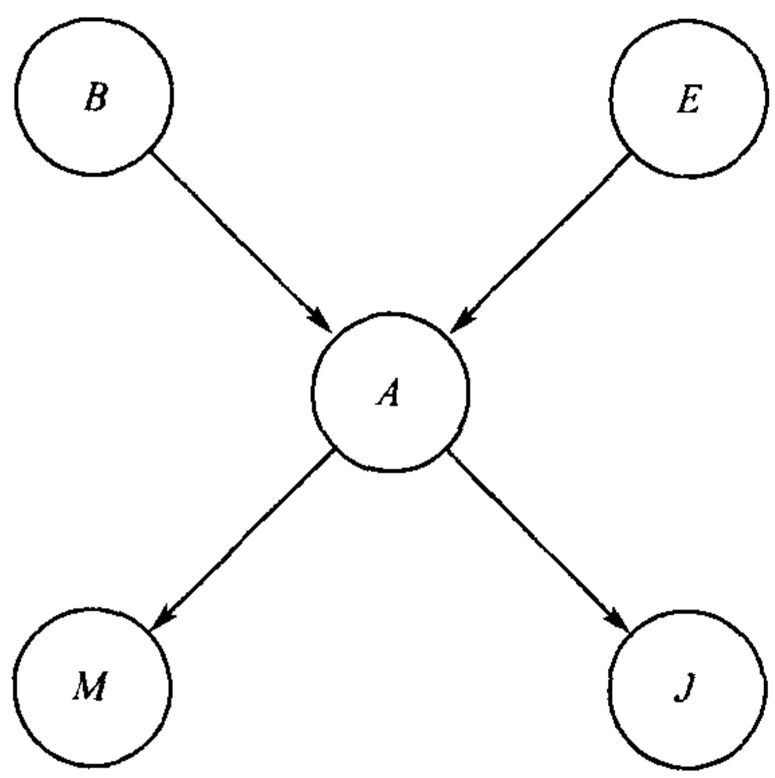


图 3.1　Alarm 网络

另一方面，如果 X 和 Y 不直接相连，那么信息需要通过其它变量才能在两者之间传递. 如果 X 和 Y 之间的所有信息通道都被阻塞 (blocked)，那么信息就无法在它们之间传递. 这时，对其中一个变量的了解不会影响对另一个变量的信度，因而 X 和 Y 相互条件独立. 那么，在什么样的条件下，X 与 Y 之间的所有信息通道都被阻塞呢?

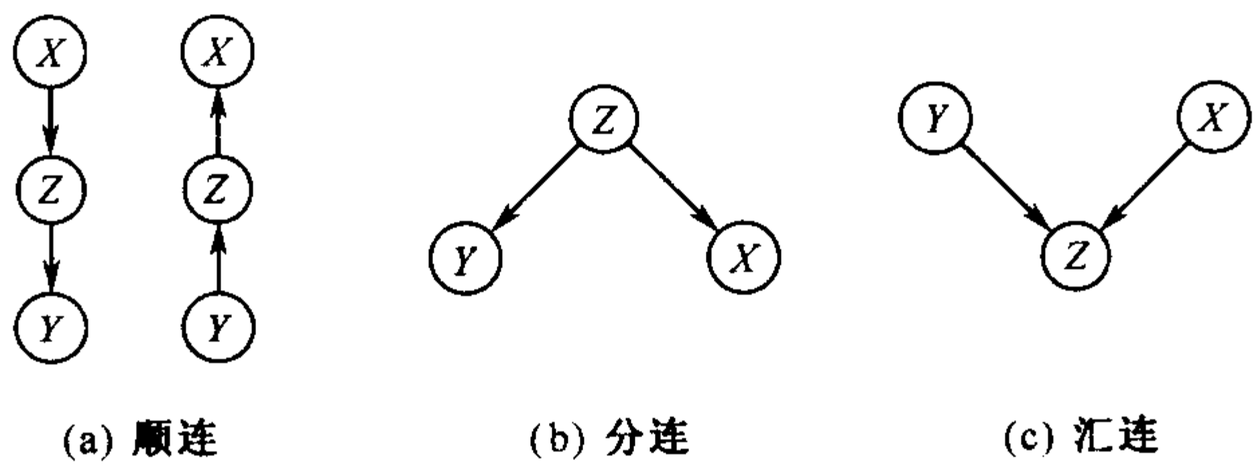
3.1.1　基本情况

考虑两个变量 X 和 Y 通过第 3 个变量 Z 间接相连这一基本情况. 它又分为 3 个子情况：顺连、分连、及汇连，如图 3.2 所示.

1. 顺连

顺连 (serial connection) 结构如图 3.2(a)所示. 这里若 Z 未知，则对 X 的

了解会影响关于 Z 的信度，进而影响关于 Y 的信度；反之亦然. 所以，此时信息可以在 X 和 Y 之间传递，它们相互关联. 另一方面，若 Z 的取值已知，则对 X 的了解就不会影响关于 Z 的信度，从而也不会影响关于 Y 的信度. 同样，对 Y 的了解也不会通过 Z 影响关于 X 的信度. 所以，此时 X 和 Y 之间的信息通道被阻塞，信息无法在两者之间传递，X 和 Y 相互条件独立.



(a) 顺连　　(b) 分连　　(c) 汇连

图 3.2　两变量 X 和 Y 通过第 3 个变量 Z 间接相连的 3 种情况

例 3.1　在图 3.1 所示的 Alarm 网络中，存在一个从盗窃（B）到警铃响（A）再到 Mary 打电话（M）的顺连结构："$B \rightarrow A \rightarrow M$". 若不知道 A 的状态，"接到 Mary 的电话"会增加对"警铃响了"的信度，从而增加对"发生了盗窃"的信度. 反之，若被告知"发生了盗窃"，则也会增加对"警铃响了"的信度，从而增加对"Mary 打了电话"的信度. 另一方面，如果早上出门前特意关闭了警铃，则知道"发生了盗窃"不会影响对"警铃响了"的信度，所以不会进而增加对"Mary 打了电话"的信度. 同理，知道"Mary 打了电话"也不会增加对"发生了盗窃"的信度，因为警铃已被关闭，所以会认为 Mary 可能是听摇滚乐太多而产生了幻听. 所以在未知 A 时，B 和 M 相互关联，而在已知 A 时，B 和 M 相互条件独立. □

2. 分连

分连（diverging connection）结构如图 3.2(b)所示. 它与顺连的情况相似：当未知 Z 时，信息可以在 X 和 Y 之间传递，它们相互关联；而当 Z 已知时，信息不能在 X 和 Y 之间传递，因而它们相互条件独立.

例 3.2　在 Alarm 网络中，警铃响（A）后 Mary（M）和 John（J）都可能打电话，即有分连结构："$M \leftarrow A \rightarrow J$". 若接到 Mary 的电话，则对"警铃响了"的信度会增大，从而也会进一步期望 John 的电话；反之亦然. 所以，未知 A 时，M 和 J 相互关联. 但如果事先知道警铃已被关掉，就不会做出这样的推理，即已知 A 后，M 和 J 条件独立. □

3. 汇连

汇连（converging connection）结构如图 3.2(c)所示，这种情况需要特别注意. 在结构上，它与分连恰恰相反. 分连代表一因多果的情况，而汇连则代表多因一果. 在信息通道性质方面，它也与分连相反：具体地说，在未知 Z 时，X 和 Y 相互独立；而在已知 Z 时，X 和 Y 却相互关联.

例 3.3 在 Alarm 网络中，盗窃（B）和地震（E）都会触响警铃（A），从而有汇连结构："$E \rightarrow A \leftarrow B$". 在 A 未知时，B 和 E 相互独立，得知"发生了地震"不会改变对"发生了盗窃"的信度，反之亦然. 但是如果知道"警铃已响"，B 和 E 却相互关联：若得知"发生了地震"，"警铃响"就有了合理的解释，从而对"发生了盗窃"的信度会降低；反过来，若得知"发生了盗窃"，"警铃响"也有了合理的解释，从而对"发生了地震"的信度也会降低. 这种同一结果的多个解释之间此消彼长的现象称为**得释**（explain away）. □

汇连的信息通道性质可以形象地理解如下：当 Z 未知时，源自 X（或 Y）的信息会从 Z"漏掉"，从而无法到达 Y（或 X）；当 Z 已知时，"漏洞"被堵上，从而信息可以在 X 和 Y 之间传递. 为了统一术语，我们称 X 和 Y 之间的信息通道在 Z 未知时被阻塞，在 Z 已知时连通.

3.1.2 一般情况

在一个贝叶斯网中，两节点 X 和 Y 之间的一条**通路**是开始于 X 结束于 Y 的一个节点序列，其中节点各异且在序列中相邻的节点在贝叶斯网中都有边将它们相连. 在例 3.4 所示的网络中，"$A \rightarrow D \rightarrow B$"和"$A \rightarrow D \rightarrow H \rightarrow K \leftarrow I \leftarrow E \leftarrow B$"是 A 与 B 之间的两条通路. 第一条称为**顺连通路**，因为其中所有边都指向同一个方向. 第二条通路不是顺连通路，因为其中有指向不同方向的边.

设 Z 为某通路上的一个节点. 如果 Z 与它前后两个节点形成一个顺连结构，则称它为一个（相对该通路的）**顺连节点**；如果 Z 与它前后两个节点形成一个分连结构，则称 Z 为一个**分连节点**；如果 Z 与它前后两个节点形成一个汇连结构，则称 Z 为一个**汇连节点**.

设 $\mathbf{Z}$ 为一节点集合，X 和 Y 是不在 $\mathbf{Z}$ 中的两个节点. 考虑 X 和 Y 之间的一条通路 α. 如果满足下面条件之一，则称 α 被 $\mathbf{Z}$ 所**阻塞**：

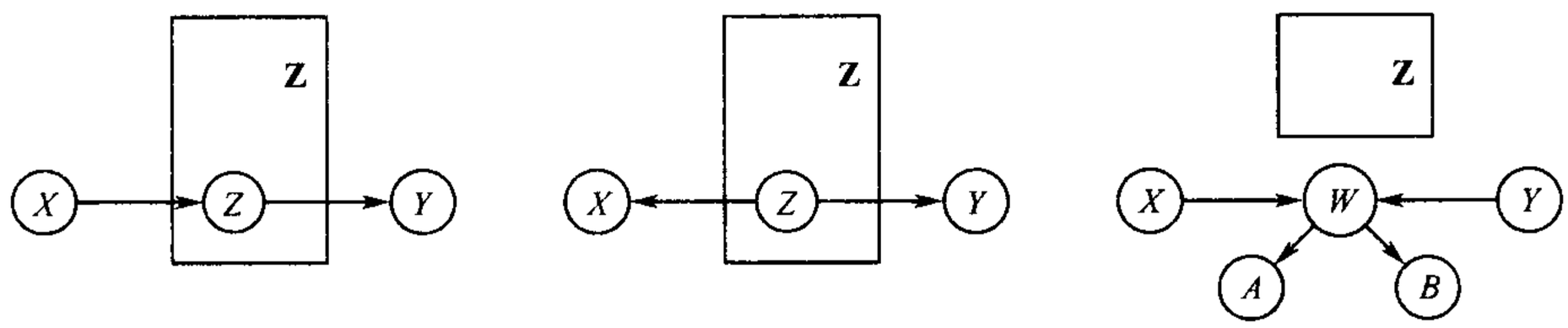
(1) α 上有一个在 $\mathbf{Z}$ 中的顺连节点；

(2) α 上有一个在 $\mathbf{Z}$ 中的分连节点；

(3) α 上有一个汇连节点 W，它和它的后代节点均不在 $\mathbf{Z}$ 中.

上述 3 种情况如图 3.3 所示.

如果通路 α 被 $\mathbf{Z}$ 所阻塞，那么当已知 $\mathbf{Z}$ 中变量的取值时，信息就不能沿着 α



(a) 顺连节点$Z\in\mathbf{Z}$　　(b) 分连节点$Z\in\mathbf{Z}$　　(c) 汇连节点W及其后代均不在$\mathbf{Z}$内

图 3.3　X 和 Y 之间的通路被 $\mathbf{Z}$ 阻塞的 3 种情况

在 X 和 Y 之间传递. 在前两种情况下，从 X 来的信息不能改变关于变量 $\mathbf{Z}$ 的信度，从而不会通过 $\mathbf{Z}$ 影响关于 Y 的信度. 而在第 3 种情况下，从 X 来的信息会从 W 处“漏掉”，从而不会到达 Y.

注意在第 3 种情况下，不仅要求 W 不在 $\mathbf{Z}$ 中，还要求 W 的所有后代节点都不在 $\mathbf{Z}$ 中. 在 Alarm 例子中，如果不知警铃是否响过，但知道 Mary 打过电话，则地震和盗窃仍然会相互关联，因为 Mary 打过电话提示了警铃可能响过.

如果 X 和 Y 之间的所有通路都被 $\mathbf{Z}$ 阻塞，那么我们就说 $\mathbf{Z}$ **有向分隔** (directed separate) X 和 Y，简称 d-**分隔** (d-separate) X 和 Y. 如果 $\mathbf{Z}$ d-分隔 X 和 Y，那么当 $\mathbf{Z}$ 中的变量全部被观测到时，信息就不能在 X 和 Y 之间传递，故 X 和 Y 相互独立. 换句话说，如果 $\mathbf{Z}$ d- 分隔 X 和 Y，那么 X 和 Y 在给定 $\mathbf{Z}$ 时条件独立.

设 $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ 是 3 个两两交空的节点集合，如果 $\mathbf{X}$ 和 $\mathbf{Y}$ 中的任意两节点 $X\in\mathbf{X}$ 和 $Y\in\mathbf{Y}$ 都被 $\mathbf{Z}$ d- 分隔，那么称 $\mathbf{Z}$ d- 分隔 $\mathbf{X}$ 和 $\mathbf{Y}$. 这时，$\mathbf{X}$ 和 $\mathbf{Y}$ 在给定 $\mathbf{Z}$ 时条件独立.

例 3.4　下面以图 3.4 所示的贝叶斯网为例，给出几个 d-分隔的例子.

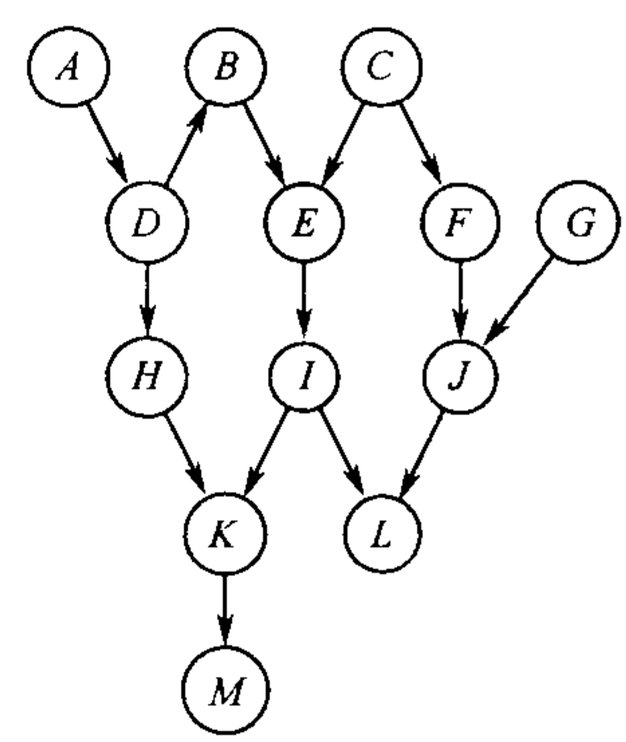


图 3.4　d-分隔示例

(1) 设 $\mathbf{Z}=\varnothing$. 先考虑节点 A 和 E，它们之间有通路“$A\rightarrow D\rightarrow B\rightarrow E$”，它未被 $\mathbf{Z}$ 阻塞，因此 A 和 E 不被 $\mathbf{Z}$ d- 分隔. 再考虑节点 A 和 C，两者之间有如下 4 条通路：

①“$A\rightarrow D\rightarrow B\rightarrow E\leftarrow C$”；

②“$A\rightarrow D\rightarrow B\rightarrow E\rightarrow I\rightarrow L\leftarrow J\leftarrow F\leftarrow C$”；

③“$A\rightarrow D\rightarrow H\rightarrow K\leftarrow I\leftarrow E\leftarrow C$”；

④“$A\rightarrow D\rightarrow H\rightarrow K\leftarrow I\rightarrow L\leftarrow J\leftarrow F\leftarrow C$”.

4 条通路中都存在汇连节点，且这些节点及其后代均不在 $\mathbf{Z}$ 中，所以 A 和 C 被 $\mathbf{Z}$ d-分隔.

(2) 设 $\mathbf{Z}=\{B, M\}$，再次考查 A 和 C 是否被 $\mathbf{Z}$ d-分隔. 这时在两者之间的 4

条通路中，前 2 条在节点 B 处有顺连结构，所以被 B 阻塞；后 2 条在 K 处有汇连结构，但 K 有一个后代节点 $M \in \mathbf{Z}$，所以这 2 条通路不被阻塞. 故 A 和 C 不被 $\mathbf{Z}$ d- 分隔.

(3) 仍设 $\mathbf{Z} = \{B, M\}$，考查节点 A 和 G 是否被 $\mathbf{Z}$ d- 分隔. 所有从 A 到 G 的通路要么经过 F 和 J 到达 G，要么经过 I，L 和 J 到达 G. 前者 在 J 处有汇连结构 "$F \rightarrow J \leftarrow G$"，且 J 及其所有后代节点都不在 $\mathbf{Z}$ 中. 后者 在 L 处有汇连结构 "$I \rightarrow L \leftarrow J$"，且 L 及其所有后代节点也都不在 $\mathbf{Z}$ 中. 因此，所有从 A 到 G 的通路都被 $\mathbf{Z}$ 阻塞，A 和 G 被 $\mathbf{Z}$ d- 分隔. □

3.2 有向分隔与条件独立

第 3.1 节引入了 d-分隔的概念，并且通过直观分析得到了 d-分隔意味着条件独立这一结论. 这一节对此进行严格证明.

3.2.1 几个引理

从一个贝叶斯网中除去一个节点 Y 即是从中删去 Y 这个节点、与 Y 所连的边以及 Y 的概率分布. 一般来说，从贝叶斯网中除去某个节点，所剩下的不一定是一个贝叶斯网. 例如在 Alarm 网络中，除去节点 B 所得到的就不是一个贝叶斯网. 它要成为贝叶斯网，A 的概率分布应该是 $P(A \mid E)$，但除去 B 后所剩下的却是 $P(A \mid B,E)$. 不过，从贝叶斯网中除去一个叶节点，所剩下的一定还是个贝叶斯网.

引理 3.1 设 $\mathcal{N}$ 是一贝叶斯网，Y 是它的一个叶节点，$\mathcal{N}'$ 是从 $\mathcal{N}$ 中除去 Y 后得到的贝叶斯网. 令 $\mathbf{X}$ 为 $\mathcal{N}'$ 中所有节点的集合，那么有

$$P_{\mathcal{N}}(\mathbf{X}) = P_{\mathcal{N}'}(\mathbf{X}),$$

即 $\mathbf{X}$ 在 $\mathcal{N}$ 中的分布函数和它在 $\mathcal{N}'$ 中的分布函数相同，或者说从 $\mathcal{N}$ 中除去 Y 不会影响 $\mathbf{X}$ 的分布函数.

证明

$$\begin{aligned}
P_{\mathcal{N}}(\mathbf{X}) &= \sum_{Y} P_{\mathcal{N}}(\mathbf{X}, Y) \\
&= \sum_{Y} \prod_{X \in \mathbf{X}} P(X \mid \pi(X)) P(Y \mid \pi(Y)) \\
&= \prod_{X \in \mathbf{X}} P(X \mid \pi(X)) \sum_{Y} P(Y \mid \pi(Y)) \\
&\quad (\text{因为 } Y \text{ 是叶节点，所以 } \forall X \in \mathbf{X}, Y \notin \pi(X)) \\
&= \prod_{X \in \mathbf{X}} P(X \mid \pi(X)) \\
&= P_{\mathcal{N}'}(\mathbf{X}).
\end{aligned}$$

引理得证. □

设 **X** 为贝叶斯网中的一个节点集合，如果其中每个节点的祖先节点都在 **X** 内，即有 $\forall X \in \mathbf{X}$, $\mathrm{an}(X) \subseteq \mathbf{X}$, 则称 **X** 是一个**祖先闭集**（ancestral set）. 对任一节点集合 **Y**, 用符号 $\mathrm{an}(\mathbf{Y})$ 代表包含 **Y** 的**最小祖先闭集**，不难看出 $\mathrm{an}(\mathbf{Y}) = \mathbf{Y} \cup \mathrm{an}(\mathbf{Y})$. 例如，在图 3.4 中，$S_1 = \{A, B, C, D, E\}$ 是一个祖先闭集，而 $S_2 = \{I, G\}$ 不是一个祖先闭集，包含 S_2 的最小祖先闭集是 $\mathrm{an}(S_2) = \{I, G, A, B, C, D, E\}$.

引理 3.2　如果 **X** 是贝叶斯网 $\mathcal{N}$ 中的一个祖先闭集，而且在 **X** 之外有节点存在，那么在 **X** 之外一定存在一个叶节点.

证明　设 Y 为 **X** 之外的一个节点. 若 Y 是叶节点，引理得证. 若否，则 Y 一定有一个子节点 Z. 因 **X** 是祖先闭集而 $Y \notin \mathbf{X}$, 必有 $Z \notin \mathbf{X}$. 重复这样的推理，最终一定可以找到一个 在 **X** 之外的叶节点. 引理得证. □

设 **X** 为贝叶斯网中的一个祖先闭集，从 $\mathcal{N}$ 中除去所有不属于 **X** 的节点而得到的也是一个贝叶斯网.

命题 3.1　设 **X** 为贝叶斯网 $\mathcal{N}$ 中的一个祖先闭集，$\mathcal{N}'$ 为从 $\mathcal{N}$ 中除去所有不属于 **X** 的节点后得到的贝叶斯网，那么有

$$P_{\mathcal{N}}(\mathbf{X}) = P_{\mathcal{N}'}(\mathbf{X}).$$

即 **X** 在 $\mathcal{N}$ 中的分布函数与它在 $\mathcal{N}'$ 中的分布函数相同.

证明　若 **X** 包含 $\mathcal{N}$ 中所有节点，命题得证. 否则，根据引理 3.2，我们一定可以找到一个叶节点 $Y \notin \mathbf{X}$. 记从 $\mathcal{N}$ 中除去 Y 所得的贝叶斯网为 $\mathcal{N}_1$. 在 $\mathcal{N}_1$ 中，**X** 仍是一个祖先闭集. 因此可以重复以上过程，直到所有不属于 **X** 的节点都被除去，最终将得到贝叶斯网 $\mathcal{N}'$. 根据引理 3.1，在此过程中 **X** 的分布函数始终保持不变，故 $P_{\mathcal{N}}(\mathbf{X}) = P_{\mathcal{N}'}(\mathbf{X})$. 命题得证. □

3.2.2　马尔可夫性

本小节的重点是证明贝叶斯网的整体马尔可夫性（定理 3.1），即 d-分隔意味着条件独立这一结论. 在给出它的证明之前，先来看一个命题.

命题 3.2　设 **X**, **Y** 和 **Z** 为贝叶斯网 $\mathcal{N}$ 中 3 个两两交空的节点集合，它们的并集为 $\mathcal{N}$ 中所有节点. 如果 **Z** d-分隔 **X** 和 **Y**, 那么 **X** 和 **Y** 在给定 **Z** 时条件独立，即

$$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}.$$

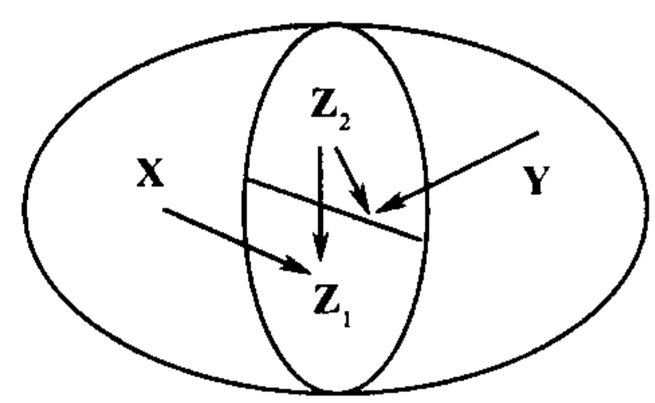


图 3.5　$\mathbf{Z}_1$ 包括 **Z** 中所有父节点属于 **X** 的节点，$\mathbf{Z}_2 = \mathbf{Z} \setminus \mathbf{Z}_1$

证明　设 $\mathbf{Z}_1$ 是 **Z** 中所有父节点属于 **X** 的节点的集合，$\mathbf{Z}_2 = \mathbf{Z} \setminus \mathbf{Z}_1$，如图 3.5 所示. 首先证明

第一个结论：对任意 $W \in \mathbf{X} \cup \mathbf{Z}_1$，$\pi(W) \subseteq \mathbf{X} \cup \mathbf{Z}$. 当 $W \in \mathbf{X}$ 时，若 $\pi(W) \nsubseteq \mathbf{X} \cup \mathbf{Z}$，即是说 W 有父节点在 $\mathbf{Y}$ 中，这与 $\mathbf{X}$ 和 $\mathbf{Y}$ 被 $\mathbf{Z}$ d-分隔的条件矛盾. 当 $W \in \mathbf{Z}_1$ 时，它必有一个父节点在 $\mathbf{X}$ 中. 若 $\pi(W) \nsubseteq \mathbf{X} \cup \mathbf{Z}$，那么它又有一个父节点在 $\mathbf{Y}$ 中，这也与 $\mathbf{X}$ 和 $\mathbf{Y}$ 被 $\mathbf{Z}$ d-分隔的条件矛盾. 所以，对任意 $W \in \mathbf{X} \cup \mathbf{Z}_1$，必有 $\pi(W) \subseteq \mathbf{X} \cup \mathbf{Z}$.

接下来证明第二个结论：对任意 $W \in \mathbf{Y} \cup \mathbf{Z}_2$，$\pi(W) \subseteq \mathbf{Y} \cup \mathbf{Z}$. 当 $W \in \mathbf{Y}$ 时，必有 $\pi(W) \subseteq \mathbf{Y} \cup \mathbf{Z}$，否则就与 $\mathbf{Z}$ d-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$ 的条件矛盾. 当 $W \in \mathbf{Z}_2$ 时，按照 $\mathbf{Z}_2$ 的定义，W 没有父节点在 $\mathbf{X}$ 中. 因此，有 $\pi(W) \subseteq \mathbf{Y} \cup \mathbf{Z}$. 所以，对任意 $W \in \mathbf{Y} \cup \mathbf{Z}_2$，$\pi(W) \subseteq \mathbf{X} \cup \mathbf{Z}$.

最后考虑所有变量的联合分布

$$
\begin{aligned}
P_{\mathcal{N}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) &= \prod_{W \in \mathbf{X} \cup \mathbf{Z} \cup \mathbf{Y}} P(W \mid \pi(W)) \\
&= \Big[\prod_{W \in \mathbf{X} \cup \mathbf{Z}_1} P(W \mid \pi(W))\Big]\Big[\prod_{W \in \mathbf{Z}_2 \cup \mathbf{Y}} P(W \mid \pi(W))\Big].
\end{aligned}
$$

根据前面的两个结论，式中第一项是 $\mathbf{X}$ 和 $\mathbf{Z}$ 的函数，记之为 $f(\mathbf{X},\mathbf{Z})$；而式中第二项是 $\mathbf{Z}$ 和 $\mathbf{Y}$ 的函数，记之为 $g(\mathbf{Z},\mathbf{Y})$. 于是，上式可改写为

$$P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) = f(\mathbf{X},\mathbf{Z})g(\mathbf{Z},\mathbf{Y}).$$

由命题 1.1，有

$$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}.$$

命题得证. □

定理 3.1（整体马尔可夫性） 设 X 和 Y 为贝叶斯网 $\mathcal{N}$ 中的两个变量，$\mathbf{Z}$ 为 $\mathcal{N}$ 中一个不包含 X 和 Y 的节点集合. 如果 $\mathbf{Z}$ d-分隔 X 和 Y，那么 X 和 Y 在给定 $\mathbf{Z}$ 时条件独立，即

$$X \perp Y \mid \mathbf{Z}.$$

证明 由于命题 3.1，可以假设 $\mathrm{an}(\{X, Y\} \cup \mathbf{Z})$ 包含 $\mathcal{N}$ 中所有节点. 设 $\mathbf{X}$ 为所有未被 $\mathbf{Z}$ 将其与 X d-分隔的节点的集合，设 $\mathbf{Y}$ 是所有既不在 $\mathbf{X}$ 也不在 $\mathbf{Z}$ 里的节点的集合. 注意 $X \in \mathbf{X}$，$Y \in \mathbf{Y}$. 因为命题 3.2，$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$. 所以一定存在函数 $f(\mathbf{X},\mathbf{Z})$ 和 $g(\mathbf{Z},\mathbf{Y})$ 使得

$$P(\mathbf{X},\mathbf{Z},\mathbf{Y}) = f(\mathbf{X},\mathbf{Z})g(\mathbf{Z},\mathbf{Y}).$$

令 $\mathbf{X}' = \mathbf{X} \setminus \{X\}$ 和 $\mathbf{Y}' = \mathbf{Y} \setminus \{Y\}$，则上式可改写为

$$P(X,\mathbf{X}',\mathbf{Z},Y,\mathbf{Y}') = f(X,\mathbf{X}',\mathbf{Z})g(\mathbf{Z},Y,\mathbf{Y}').$$

于是

$$P(X, Y, \mathbf{Z}) = \sum_{\mathbf{X}', \mathbf{Y}'} P(X, \mathbf{X}', \mathbf{Z}, Y, \mathbf{Y}')$$

$$
\begin{aligned}
&= \sum_{\mathbf{X}', \mathbf{Y}'} f(X, \mathbf{X}', \mathbf{Z}) g(\mathbf{Z}, Y, \mathbf{Y}') \\
&= \Big[\sum_{\mathbf{X}'} f(X, \mathbf{X}', \mathbf{Z})\Big]\Big[\sum_{\mathbf{Y}'} g(\mathbf{Z}, Y, \mathbf{Y}')\Big] \\
&= f'(X, \mathbf{Z}) g'(\mathbf{Z}, Y).
\end{aligned}
$$

所以

$$X \perp Y \mid \mathbf{Z}.$$

定理得证. □

定理 3.1 说的是，在贝叶斯网中，d-分隔意味着条件独立. d-分隔是图论的概念，而条件独立是概率论的概念，所以定理 3.1 揭示了贝叶斯网的图论侧面和它的概率论侧面之间的关系.

在贝叶斯网中，一个节点 X 的**马尔可夫边界** (Markov boundary)[①] 包括其父节点、子节点以及子节点的父节点，即 $\mathrm{mb}(X) = \pi(X) \cup \mathrm{ch}(X) \cup \bigcup_{Y \in \mathrm{ch}(X)} \pi(Y)$. 在图 3.4 中，节点 I 的马尔可夫边界为 $\{E, H, J, K, L\}$. 下面是定理 3.1 的两个推论.

推论 3.1　在一个贝叶斯网中，给定变量 X 的马尔可夫边界 $\mathrm{mb}(X)$，则 X 条件独立于网络中所有其它变量.

证明　设 $\mathbf{Y}$ 是网络中所有其它变量的集合. 如图 3.6 所示，考虑从 X 出发到 $\mathbf{Y}$ 中某一节点的一条通路 α，设 γ 是 α 离开 $\mathrm{mb}(X)$ 之前的最后一个节点. 根据 $\mathrm{mb}(X)$ 的定义，α 在 γ 处的连接方式一定是分连或顺连. 因此，α 被 $\mathrm{mb}(X)$ 所阻塞. 由于 α 的任意性，$\mathrm{mb}(X)$ d-分隔 X 和 $\mathbf{Y}$. 所以给定 $\mathrm{mb}(X)$，X 与 $\mathbf{Y}$ 条件独立. 推论得证. □

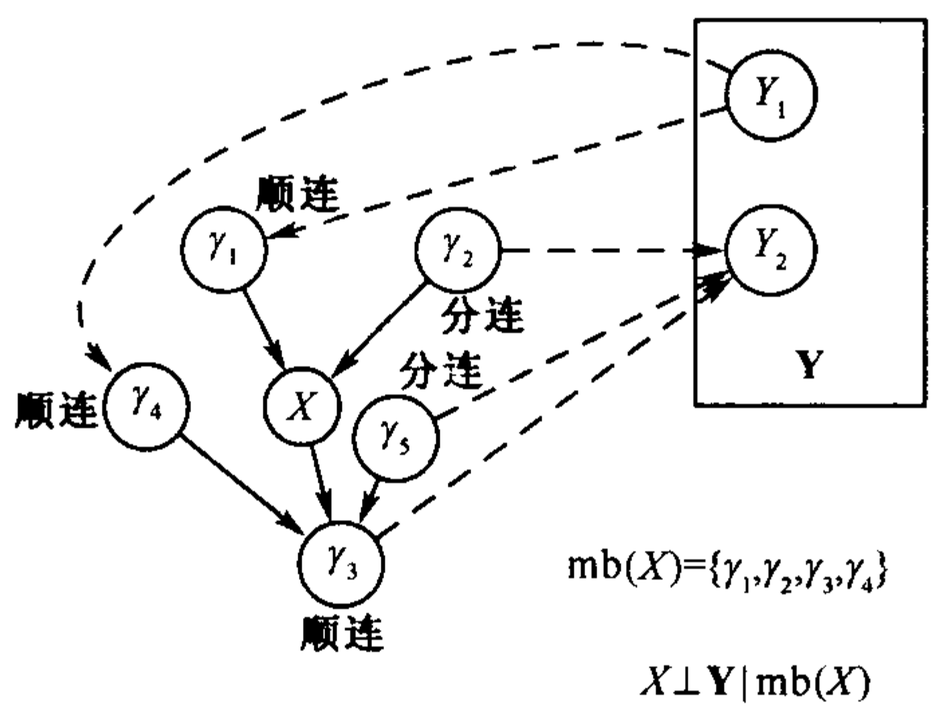


图 3.6　给定马尔可夫边界 mb（X），X 条件独立于所有其它节点 $\mathbf{Y}$

① 有时也称为 Markov blanket.

推论 3.2（局部马尔可夫性） 在一个贝叶斯网中，给定变量 X 的父节点 $\pi(X)$，则 X 条件独立于它的所有非后代节点，即

$$X \perp [\mathrm{nd}(X) \setminus \pi(X)] \mid \pi(X).$$

证明 因为定理 3.1，只需证明 $\pi(X)$ d-分隔 X 和 $\mathrm{nd}(X) \setminus \pi(X)$. 考虑从 X 到 $\mathrm{nd}(X) \setminus \pi(X)$ 中某一节点 Y 的一通路 α，用 γ 记其中与 X 相邻的节点. 如图 3.7 所示，有两种情况：① $\gamma \in \pi(X)$，此时 α 在 X 处的连接情况一定是分连或顺连，因此 α 被 $\pi(X)$ 所阻塞；② $\gamma \notin \pi(X)$，此时，由于 Y 不是 X 的后代节点，所以 α 中一定存在一个汇连节点 γ_0，且 $\gamma_0 \notin \pi(X)$. 因此 α 还是被 $\pi(X)$ 所阻塞. 由 Y 和 γ 的任意性，$\pi(X)$ d-分隔 X 和 $\mathrm{nd}(X)$. 推论得证. □

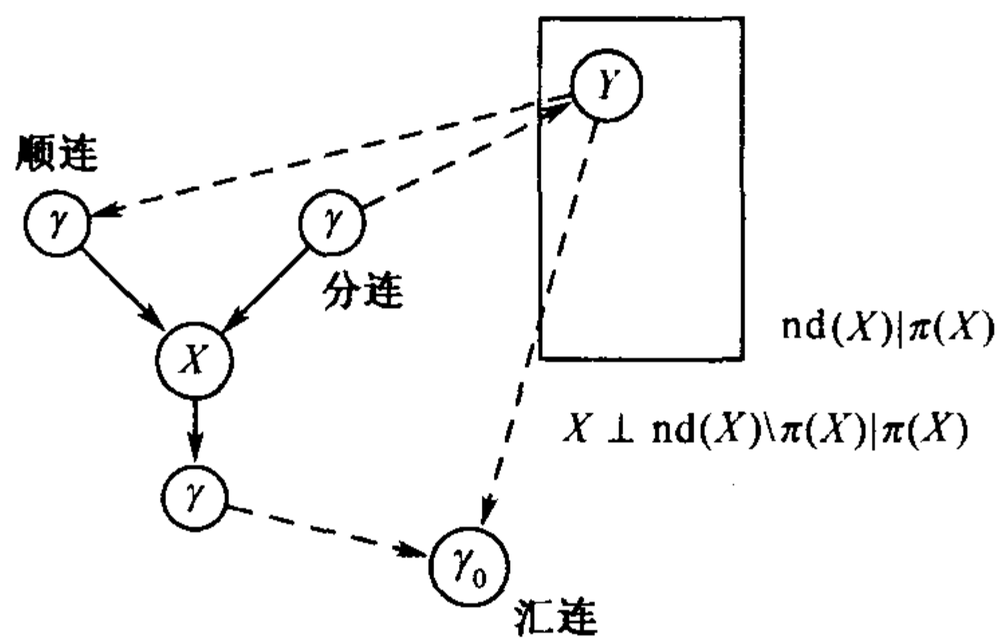


图 3.7 给定父节点 $\pi(X)$，X 条件独立于 $\mathrm{nd}(X)\setminus\pi(X)$

第 2.4.2 节曾提到，当利用因果关系建立贝叶斯网网络结构时，实际上是在基于因果关系进行条件独立假设. 现在可以把这个假设说清楚了：

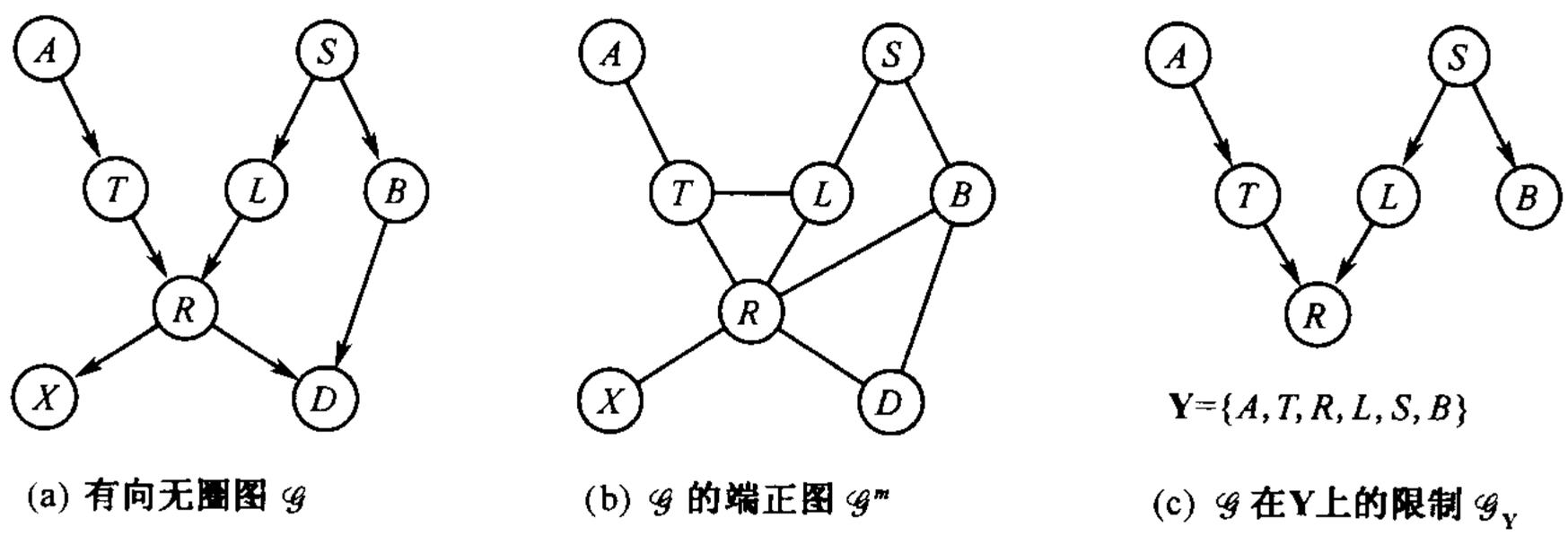
因果马尔可夫假设：给定一个变量 X 的直接原因（父节点），该变量条件独立于所有那些不是它的直接或间接结果（非后代节点）的变量.

3.3 有向分隔与无向分隔

本节讨论有向无圈图 中的 d-分隔与无向图中的无向分隔之间的关系，这一关系有助于判定 d-分隔.

设 $\mathscr{G}$ 为一有向无圈图 . 如果将 $\mathscr{G}$ 中每个节点的不同父节点**结合**，即在它们之间加一条边，然后去掉所有边的方向，所得到的无向图称为 $\mathscr{G}$ 的**端正图**（moral graph），记之为 $\mathscr{G}^m$. 从 $\mathscr{G}$ 到 $\mathscr{G}^m$ 的过程称为**端正化**（moralization）. 考虑图 3.8(a)所示的有向无圈图 . 要把它端正化，首先要将图中节点 R 的两个父节点 T 和 L 结合，把 D 的 两个父节点 R 和 B 结合，然后去掉所有边的方向. 其结果如图 3.8(b)所示.

设 **Y** 是有向无圈图 $\mathscr{G}$ 中的一个节点集合. $\mathscr{G}$ 在 **Y** 上的**限制**，记为 $\mathscr{G}_{\mathbf{Y}}$，是从 $\mathscr{G}$ 中除去所有不属于 **Y** 的节点及其相连的边而得到的图. 例如，图 3.8(a)所示的 $\mathscr{G}$ 在集合 $\mathbf{Y}=\{A, T, R, L, S, B\}$ 上的限制如图 3.8(c)所示.



(a) 有向无圈图 $\mathscr{G}$　(b) $\mathscr{G}$ 的端正图 $\mathscr{G}^m$　(c) $\mathscr{G}$ 在Y上的限制 $\mathscr{G}_{\mathbf{Y}}$

图 3.8　有向无圈图、端正图及图限制

在一个无向图 $\mathscr{G}$ 中，节点 X 和 Y 之间的一条**通路**指的是从 X 开始到 Y 结束的一个节点序列，其中节点各异且在序列中相邻的节点在 $\mathscr{G}$ 中也相邻. 设 **X**，**Y**，**Z** 是 $\mathscr{G}$ 中的 3 个两两交空的节点集合. 如果从 $\mathscr{G}$ 中 除去 **Z** 中的节点后，**X** 和 **Y** 之间没有通路存在，则称 **Z 无向分隔**（undirected separate）**X** 和 **Y**，简称 **Z** u-**分隔** **X** 和 **Y**. 在图 3.8(b)中，$\{R, L\}$ u-分隔 A 和 D，但 $\{R\}$ 则不 u-分隔 A 和 D. 容易看到，**Z** 在 $\mathscr{G}$ 中 u-分隔 **X** 和 **Y** 当且仅当 **X** 和 **Y** 之间的任何通路都经过 **Z**.

定理 3.2（有向分隔与无向分隔）　设 **X**,**Y**,**Z** 是有向无圈图 $\mathscr{G}$ 中 3 个两两交空的节点集合. 用 $(\mathscr{G}_{\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})})^m$ 表示先把 $\mathscr{G}$ 限制在 $\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})$ 上，再将其结果端正化而得到的无向图. 那么，集合 **Z** 在 $\mathscr{G}$ 中 d-分隔 **X** 和 **Y** 的充分必要条件是它在 $(\mathscr{G}_{\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})})^m$ 中 u-分隔 **X** 和 **Y**.

证明　下面把 $(\mathscr{G}_{\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})})^m$ 简记为 $\mathscr{G}'$. 先证明充分性：如果 **Z** 在 $\mathscr{G}$ 中不 d-分隔 **X** 和 **Y**，那么它在 $\mathscr{G}'$ 中不 u-分隔 **X** 和 **Y**.

由于 **Z** 在 $\mathscr{G}$ 中不 d-分隔 **X** 和 **Y**，那么一定存在 $X\in\mathbf{X}$ 和 $Y\in\mathbf{Y}$，使得 X 和 Y 之间有一条不被 **Z** 阻塞的通路 α. 设 α' 是从 α 中删除所有汇连节点而得到的节点序列. 下面证明 α' 是一条在 $\mathscr{G}'$ 中连接 X 和 Y 而且不经过 **Z** 的通路.

先证明 α' 是 $\mathscr{G}'$ 中的一条通路. 设 **C** 为 α 中所有汇连节点的集合. 考虑 α' 中的任一节点 A，有 $A\in\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{C})$. 又因为 α 未被 **Z** 阻塞，所以必有 $\mathbf{C}\subseteq\mathrm{an}(\mathbf{Z})$，因此 $A\in\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})$，即 α' 中的所有节点都在 $\mathscr{G}'$ 中. 考虑 α' 中两相邻节点 A 和 B：若在 $\mathscr{G}$ 中它们之间有一条边，那么在 $\mathscr{G}'$ 中同样有；若在 $\mathscr{G}$ 中它们之间没有边，那么它们在 α 中一定位于某个汇连节点 C 的左右. 前面已提到，$C\in\mathrm{an}(\mathbf{Z})$，所以在 $\mathscr{G}_{\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})}$ 中有边“$A\rightarrow C$”和“$C\leftarrow B$”，从而在端正化过程中会在 A 和 B 之间加一条边. 因此 α' 确实是 $\mathscr{G}'$ 中一条连接 X 和 Y 的通路.

另一方面，由 α' 的构造方法可知，其中任一节点 A 在 α 中一定不是汇连节点. 又因为 $\mathbf{Z}$ 不阻塞 α，所以 A 一定不在 $\mathbf{Z}$ 中，即 α' 不经过 $\mathbf{Z}$. 所以，α' 是 $\mathscr{G}'$ 中一条连接 X 和 Y 且不经过 $\mathbf{Z}$ 的通路，故 $\mathbf{Z}$ 在 $\mathscr{G}'$ 中不 u-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$.

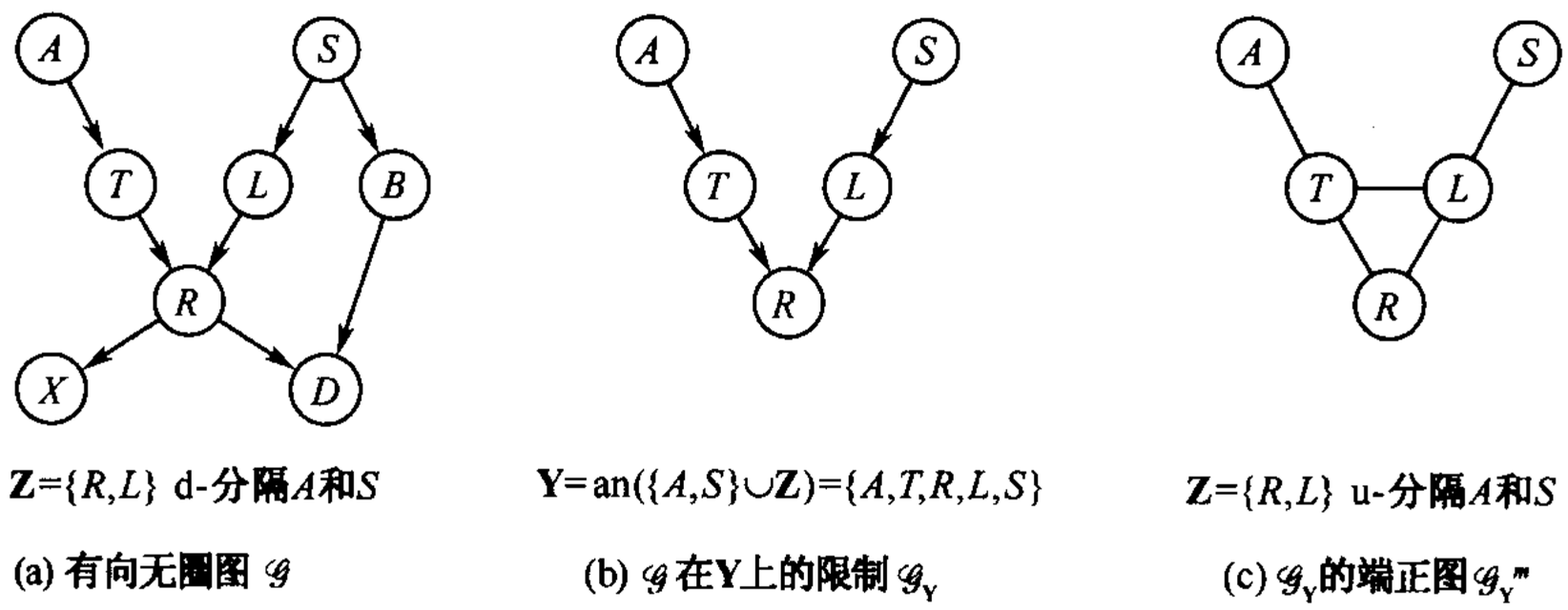
接下来证明必要性：如果 $\mathbf{Z}$ 在 $\mathscr{G}'$ 中不 u-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$，那么它在 $\mathscr{G}$ 中不 d-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$. 为此，我们先定义有向原通路的概念. 对 $\mathscr{G}'$ 中任一连接 $\mathbf{X}$ 和 $\mathbf{Y}$ 的通路 α'，可在其基础上按如下方法构造一个节点序列 α：考虑 α' 中的两个相邻节点 A 和 B，若在 $\mathscr{G}$ 中它们是邻居，则不在它们之间加入新节点；若在 $\mathscr{G}$ 中它们不是邻居，那么必定存在节点 $C \in \mathrm{an}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z})$，使得“$A \rightarrow C$”和“$C \leftarrow B$”是 $\mathscr{G}_{\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})}$ 中的边，这时将 C 加到 A 和 B 之间. 显然 α 是 $\mathscr{G}$ 中的一条连接 $\mathbf{X}$ 和 $\mathbf{Y}$ 的通路，我们称它为 α' 在 $\mathscr{G}$ 中的一条**有向原通路**. 注意，α' 在 $\mathscr{G}$ 中的有向原通路可能不止一条，但所有这些有向原通路所包含的汇连节点的个数是一样的.

由于 $\mathbf{X}$ 和 $\mathbf{Y}$ 在 $\mathscr{G}'$ 中不被 $\mathbf{Z}$ u-分隔，那么它们之间一定存在不经过 $\mathbf{Z}$ 的通路. 设 α' 是所有这些通路中其有向原通路包含汇连节点个数最少的一个，设 α 是 α' 一条有向原通路. 下面证明 α 在 $\mathscr{G}$ 中不被 $\mathbf{Z}$ 阻塞.

考虑 α 中任一节点 D：若 D 在 α 中不是汇连节点，则它一定来自 α'，因而不在 $\mathbf{Z}$ 中；若 D 是 α 中的一汇连节点，如果可以证明 $D \in \mathrm{an}(\mathbf{Z})$，那么 $\mathbf{Z}$ 将不阻塞 α. 下面证明 $D \in \mathrm{an}(\mathbf{Z})$.

由有向原通路的构造方法可知，$D \in \mathrm{an}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z})$. 若 $D \notin \mathrm{an}(\mathbf{Z})$，则必有 $D \in \mathrm{an}(\mathbf{X})$ 或 $D \in \mathrm{an}(\mathbf{Y})$. 不失一般性，设 $D \in \mathrm{an}(\mathbf{Y})$，则在 $\mathscr{G}_{\mathrm{an}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})}$ 中一定存在一条从 D 到 $\mathbf{Y}$ 且不过 $\mathbf{Z}$ 的顺连通路 α_1. 设 α 在 D 处的汇连结构为“$A \rightarrow D \leftarrow B$”，其中 A 位于靠近 X 一侧. 从通路 α' 中取出从 $\mathbf{X}$ 到 A 的子序列，然后在其后附上 α_1 中的节点，就得到 $\mathscr{G}'$ 中的另外一条 $\mathbf{X}$ 和 $\mathbf{Y}$ 之间的不经过 $\mathbf{Z}$ 通路. 它的有向原通路所包含的汇连节点个数比 α 的少，这与 α' 的定义相矛盾. 所以，必有 $D \in \mathrm{an}(\mathbf{Z})$. 定理得证. □

下面看一个利用 u-分隔关系来判定 d-分隔的例子.



$\mathbf{Z}=\{R,L\}$ d-分隔A和S　　$\mathbf{Y}=\mathrm{an}(\{A,S\}\cup\mathbf{Z})=\{A,T,R,L,S\}$　　$\mathbf{Z}=\{R,L\}$ u-分隔A和S

(a) 有向无圈图 $\mathscr{G}$　　(b) $\mathscr{G}$ 在$\mathbf{Y}$上的限制 $\mathscr{G}_{\mathbf{Y}}$　　(c) $\mathscr{G}_{\mathbf{Y}}$的端正图$\mathscr{G}_{\mathbf{Y}}^{m}$

图 3.9　利用 u-分隔判定 d-分隔示例

例 3.5　考虑如图 3.9(a)所示的有向无圈图，设 $\mathbf{Z}=\{R, L\}$，$\mathbf{Z}$ 是否 d-分隔 A 和 S？为了回答这个问题，我们①先求得 $\mathbf{Y}=\mathrm{an}(\{A, S\}\cup\mathbf{Z})=\{A, T, R, L, S\}$；②将 $\mathscr{G}$ 限制在 $\{A, T, R, L, S\}$ 上，得到 $\mathscr{G}_{\mathbf{Y}}$，如图 3.9(b)所示；③将 $\mathscr{G}_{\mathbf{Y}}$ 端正化，得到 $(\mathscr{G}_{\mathbf{Y}})^m$，如图 3.9(c)所示. 因为在 $(\mathscr{G}_{\mathbf{Y}})^m$ 中 $\mathbf{Z}$ u-分隔 A 和 S，所以在 $\mathscr{G}$ 中 $\mathbf{Z}$ d-分隔 A 和 S. 根据这几个图还可以看到，如果 $\mathbf{Z}$ 仅包含节点 R，那么它不 d-分隔 A 和 S，因为这时 $\mathbf{Y}$ 和 $(\mathscr{G}_{\mathbf{Y}})^m$ 都没有变，但是在 $(\mathscr{G}_{\mathbf{Y}})^m$ 中节点 R 却不 u-分隔 A 和 S.　□

3.4　有向无圈图与联合概率分布

第 3.2 节讨论了贝叶斯网的图论侧面（有向无圈图中的 d-分隔）和它的概率论侧面（联合概率分布中的条件独立）之间的关系. 在本节中，我们更一般地讨论有向无圈图和联合概率分布之间的关系：一方面用有向无圈图来刻画联合概率的某些性质，另一方面则相对于联合概率分布来定义一些图论概念.

设 $\mathbf{V}$ 为一个随机变量的集合，P 是 $\mathbf{V}$ 的一个联合概率分布，$\mathscr{G}$ 是以 $\mathbf{V}$ 中变量为节点的一个有向无圈图. 设 $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ 为 $\mathbf{V}$ 的 3 个两两交空的子集合. 引入如下的两个记号：

(1) $\mathbf{X}\perp_P\mathbf{Y}\mid\mathbf{Z}$：表示在 P 中，给定 $\mathbf{Z}$，$\mathbf{X}$ 与 $\mathbf{Y}$ 相互条件独立；

(2) $S_{\mathscr{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$：表示在 $\mathscr{G}$ 中，$\mathbf{Z}$ d-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$.

利用有向无圈图 $\mathscr{G}$，可以定义如下 3 个关于联合概率分布 P 的性质：

(1) P 是 $\mathscr{G}$-**可分解的**，如果存在一个以 $\mathscr{G}$ 为结构的贝叶斯网，它所表示的联合概率分布正好是 P；

(2) P **具有整体** $\mathscr{G}$-**马尔可夫性**，如果

$$S_{\mathscr{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\Rightarrow\mathbf{X}\perp_P\mathbf{Y}\mid\mathbf{Z},$$

即 $\mathscr{G}$ 中的 d-分隔意味着 P 中的条件独立.

(3) P **具有局部** $\mathscr{G}$-**马尔可夫性**，如果对于 $\mathscr{G}$ 中任一节点 X，

$$X\perp_P\mathrm{nd}_{\mathscr{G}}(X)\mid\pi_{\mathscr{G}}(X),$$

即在 P 中，X 在给定它在 $\mathscr{G}$ 中的父节点时，条件独立于它在 $\mathscr{G}$ 中的所有非后代节点.

下面的定理揭示了上述 3 个性质之间的关系.

定理 3.3　假设 P 为一个变量集合 $\mathbf{V}$ 的联合概率分布，$\mathscr{G}$ 为定义于 $\mathbf{V}$ 上的一个有向无圈图，则以下 3 个陈述是等价的：

(1) P 是 $\mathscr{G}$-**可分解的**；

(2) P 具有**整体** $\mathscr{G}$-**马尔可夫性**；

(3) P 具有**局部** $\mathscr{G}$-**马尔可夫性**.

证明 根据定理 3.1，(1)⇒(2). 而推论 3.2 的证明显示，(2)⇒(3). 下面用归纳法证明 (3)⇒(1).

当只有一个节点时，这显然成立. 假设它对 $n-1$ 个节点的情况也成立，考虑 n 个节点的情况：设 X 为 $\mathscr{G}$ 中的一个叶节点，$\mathbf{V}' = \mathbf{V} \setminus \{X\}$. 根据 (3)，给定 $\pi_{\mathscr{G}}(X)$，X 条件独立于所有其它节点，所以有

$$P(\mathbf{V}) = P(\mathbf{V}')P(X \mid \mathbf{V}') = P(\mathbf{V}')P(X \mid \pi_{\mathscr{G}}(X)).$$

设 $\mathscr{G}'$ 为从 $\mathscr{G}$ 中移去 X 所得的图，则 $P(\mathbf{V}')$ 具有局部 $\mathscr{G}'$-马尔可夫性. 因为 $\mathbf{V}'$ 有 $n-1$ 个节点，所以根据归纳假设，$P(\mathbf{V}')$ 是 $\mathscr{G}'$-可分解的. 因此 $P(\mathbf{V})$ 是 $\mathscr{G}$-可分解的. 定理得证. □

相对于联合分布 P，可以定义如下两个关于有向无圈图 $\mathscr{G}$ 的性质：

(1) $\mathscr{G}$ 是 P 的**独立图** (independence map)，简称 I-map，如果

$$S_{\mathscr{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \Rightarrow \mathbf{X} \perp_P \mathbf{Y} \mid \mathbf{Z};$$

(2) $\mathscr{G}$ 是 P 的**依赖图** (dependence map)，简称 D-map，如果

$$\mathbf{X} \perp_P \mathbf{Y} \mid \mathbf{Z} \Rightarrow S_{\mathscr{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) .$$

若 $\mathscr{G}$ 既是 P 的独立图，也是 P 的依赖图，则称它是 P 的**完美图** (perfect map)，简称 P-map. 独立图表示 P 中的独立关系，依赖图表示 P 中的依赖关系. 独立图所蕴含的条件独立关系是 P 中所有条件独立关系的一个子集，而后者又是依赖图中所蕴含的条件独立关系的一个子集. 因此，在独立图中加一条边，所得到的仍然是独立图①. 类似地，在依赖图中去掉一条边，所得到的仍然是依赖图.

一个独立图被称为是**最小独立图** (minimal I-map)，如果去掉其中的一条边后它不再是独立图. 在第 2.4.1 节所描述的确定贝叶斯网的结构的过程中，如果在第(4)步选择最小的 $\pi(X_i)$，使得节点 X_i 在给定 $\pi(X_i)$ 后与其它节点条件独立，那么所得到的贝叶斯网是一个最小独立图.

3.5 文献介绍

有向分隔准则最早是 Verma 在一份技术报告(Verma，1987)中提出的，1988 年首次发表 (Verma and Pearl，1988). 第 3.2 节的证明与别的著作不同，其思想源自文献(Zhang et al.，1994). 第 3.3 和 3.4 两节的内容取材于文献 (Cowell et al.，1999) 的第 5 章. 在探索概率独立与图分隔的关系的过程中，学者们还对条件独立关系进行了公理化研究，详情参见文献 (Pearl，1988) 的第 3 章.

① 注意，$\mathscr{G}$ 是 P 的独立图意味着 P 是 $\mathscr{G}$-可分解的.

第二部分

贝叶斯网推理

第 4 章　贝叶斯网与概率推理

在第 2 章我们已经看到，基于联合分布进行概率推理的复杂度相对于变量个数成指数增长，因此在变量众多时是不可行的. 贝叶斯网利用变量间的条件独立关系对联合分布进行分解，减少模型的参数个数，从而使知识的获取与表达得以简化. 在这一章里，我们讨论如何利用联合分布的分解来简化推理，并且给出最基本的贝叶斯网推理算法，即变量消元法.

4.1　推理问题

推理（inference）是通过计算回答**查询**（query）的过程. 贝叶斯网中的推理问题有三大类：后验概率问题、最大后验假设问题以及最大可能解释问题.

4.1.1　后验概率问题

后验概率问题指的是已知贝叶斯网中某些变量的取值，计算另外一些变量的后验概率分布的问题. 例如，在 Alarm 网络中，若接到 Mary 的电话说“警铃响了”，这时自然会想知道“发生了盗窃”的概率是多少，即计算 $P(B=y \mid M=y)$. 在此类问题中，已知变量通常称为**证据变量**（evidence variables），记为 $\mathbf{E}$，它们的取值记为 $\mathbf{e}$；需要计算其后验概率分布的变量称为**查询变量**（query variables），记为 $\mathbf{Q}$；需要计算的后验分布为 $P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$.

人们常说的**概率推理**指的就是后验概率问题. 根据证据变量和查询变量所扮演的因果角色的不同，概率推理有以下 4 种不同类型.

（1）从结果到原因的**诊断推理**（diagnostic inference）. 例如，已知 Mary 打电话（$M=y$），计算发生盗窃的概率 $P(B=y \mid M=y)$.

（2）从原因到结果的**预测推理**（predictive inference）. 例如，已知发生了入室盗窃（$B=y$），计算 Mary 打电话的概率 $P(M=y \mid B=y)$.

（3）在同一结果的不同原因之间的**原因关联推理**（intercausal inference）. 例如，地震和入室盗窃都是导致警铃响的原因. 已知警铃响后（$A=y$），对发生盗窃的信度为 $P(B=y \mid A=y)$. 如果接着又获知发生了地震（$E=y$），则对盗窃的信度将变为 $P(B=y \mid A=y, E=y)$. 这里，由于警铃响了已有了一个可能的解释，即地震，所以此长彼消，对发生盗窃的信度将会减小，即 $P(B=y \mid A=y, E=y) < P(B=y \mid A=y)$.

(4) 包含多种上述类型的**混合推理** (mixed inference). 例如，已知 John 打电话 ($J = y$) 和发生了地震 ($E = y$)，计算警铃响了的概率 $P(A = y \mid J = y, E = y)$. 这里既有诊断推理，又有预测推理.

在不同的应用中，可能会遇到不同类型的后验概率问题，但它们都可以用同一方法来处理①.

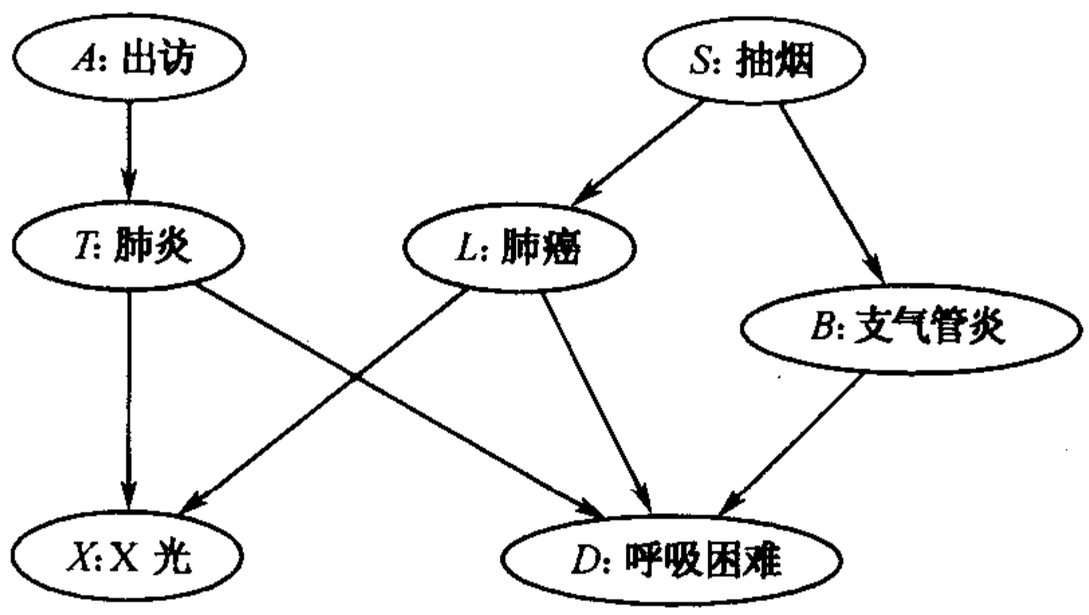
4.1.2 最大后验假设问题

已知证据 $\mathbf{E} = \mathbf{e}$，有时会对一些变量的后验概率最大的状态组合感兴趣，这些变量称为**假设变量** (hypothesis variables)，记之为 $\mathbf{H}$. $\mathbf{H}$ 的一个状态组合称为一个**假设** (hypothesis)，记之为 $\mathbf{h}$. 在所有可能的假设中，想找出后验概率最大的②那个假设 $\mathbf{h}^*$，即

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} P(\mathbf{H} = \mathbf{h} \mid \mathbf{E} = \mathbf{e}).$$

这就是**最大后验假设** (maximum a posteriori hypothesis) 问题，简称 **MAP** 问题.

例 4.1 在图 4.1 所示的肺病诊断贝叶斯网中，假设病人的 X 光胸透结果有阴影，但却没有呼吸困难的症状，即观测证据为 $\{X = y, D = n\}$. 问病人患哪个或哪些疾病的可能性最大? 有如下 8 种可能的组合：



证据: $\{X=y,D=n\}$
假设变量 H: $\{T,L,B\}$

图 4.1 肺病诊断贝叶斯网中的最大后验假设问题

(1) $T=y$, $L=y$, $B=y$;

(2) $T=y$, $L=y$, $B=n$;

---

① 这与逻辑推理不同，那里诊断推理用的是**反绎** (abduction)，而预测推理用的是**演绎** (deduction).

② 为方便计，我们假设后验概率最大的假设只有一个. 一般而言，可能存在多个概率相等的最大后验假设 $\mathbf{h}^*$.

(3) $T=y$, $L=n$, $B=y$;

(4) $T=y$, $L=n$, $B=n$;

(5) $T=n$, $L=y$, $B=y$;

(6) $T=n$, $L=y$, $B=n$;

(7) $T=n$, $L=n$, $B=y$;

(8) $T=n$, $L=n$, $B=n$.

我们感兴趣的是后验概率最大的那个组合. 这是一个最大后验假设问题，假设变量为 $\{T, L, B\}$. □

4.1.3 最大可能解释问题

在贝叶斯网中，证据 $\mathbf{E}=\mathbf{e}$ 的一个**解释**（explanation）指的是网络中全部变量的一个与 $\mathbf{E}=\mathbf{e}$ 相一致的状态组合. 往往有时最关心概率最大的那个解释，即**最大可能解释**（most probable explanation），简称 MPE ①. 求最大可能解释的 MPE 问题可视为 MAP 问题的一个特例，即 MAP 中的假设变量 $\mathbf{H}$ 包含了网络中的所有非证据变量.

例 4.2　在例 4.1 中，假设证据仍然是 $\{X=y, D=n\}$. 对它的一个解释为

$$\{A=y, S=y, T=y, L=y, B=y, X=y, D=n\}.$$

这样的解释共有 $2^5=32$ 个. MPE 就是要找出概率最大的那一个解释. □

本章主要讨论后验概率问题，只在最后简短论及 MPE 和 MAP 问题.

4.2 变量消元算法

本节揭示为什么联合分布的分解可以降低概率推理复杂性的原理，并给出一个计算后验概率 $P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$ 的基本的算法，即**变量消元法**（variable elimination).

4.2.1 概率分布的分解与推理复杂度

以图 4.2 中的贝叶斯网为例，考虑计算 $P(D)$，有

$$\begin{aligned} P(D) &= \sum_{A,B,C} P(A, B, C, D) \\ &= \sum_{A,B,C} P(A)P(B \mid A)P(C \mid B)P(D \mid C). \end{aligned} \tag{4.1}$$

假设所有变量均为二值，而且乘法按顺序进行. 式(4.1)的运算复杂度如下：$P(A)$ 与 $P(B \mid A)$ 相乘需要做 4 次数字乘法，其结果与 $P(C \mid B)$ 相乘需要做 8 次

① 同 MAP 一样，我们假设最大可能解释是唯一的.

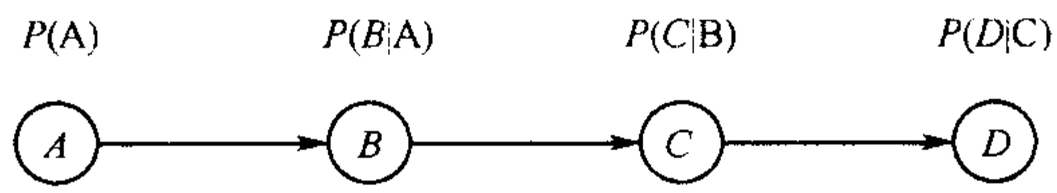


图 4.2 一个链状贝叶斯网

数字乘法，它的结果再与 $P(D \mid C)$ 相乘需要做 16 次数字乘法. 所以，总共需要做 28 次数字乘法. 另外，所需的数字加法共 14 次.

为了利用联合分布的分解来降低推理的计算复杂度，我们注意到在式（4.1）右边的 4 个因子中，只有 $P(A)$ 和 $P(B \mid A)$ 与变量 A 相关，而与 B 相关的因子只有 $P(C \mid B)$ 和 $P(B \mid A)$. 因此，有

$$P(D) = \sum_C P(D \mid C) \sum_B P(C \mid B) \sum_A P(A) P(B \mid A). \tag{4.2}$$

式（4.2）的计算复杂度如下：$P(A)$ 与 $P(B \mid A)$ 相乘需要做 4 次数字乘法，然后边缘化消去 A 需要做 2 次数字加法；这一步的结果与 $P(C \mid B)$ 相乘需要做 4 次数字乘法，然后消去 B 需要做 2 次数字加法；它的结果再与 $P(D \mid C)$ 相乘需要做 4 次数字乘法，然后消去 C 需要做 2 次数字加法. 所以，数字乘法的总次数为 $4+4+4=12$，而数字加法的总次数为 $2+2+2=6$. 这比式（4.1）的复杂度要低.

联合分布的分解之所以能够降低推理的复杂度，主要是因为它使得运算可以局部化. 在式（4.1）中，消去 A 涉及所有的变量：A，B，C 和 D，因此复杂度高. 而在式（4.2）中，消去 A 只涉及它自身以及与它直接相连的变量 B，因此复杂度低. 在上面的例子中，运算局部化节省了大约一半的运算量. 在变量众多的网络中，节省可能是指数级的. 下一节将把式（4.2）所蕴涵的原理推广为一个算法. 这里我们先指出，式（4.2）本身可以改写为如下的算法形式：

（1）设 $\mathscr{F}$ 是贝叶斯网中所有概率分布的集合，即

$$\mathscr{F} = \{\underline{P(A)}, \underline{P(B \mid A)}, P(C \mid B), P(D \mid C)\};$$

（2）从 $\mathscr{F}$ 中删去所有含变量 A 的函数 $P(A)$ 和 $P(B \mid A)$，并生成一个新函数

$$\psi_1(B) = \sum_A P(A) P(B \mid A),$$

然后将其加入 $\mathscr{F}$，得到

$$\mathscr{F} = \{\underline{\psi_1(B)}, \underline{P(C \mid B)}, P(D \mid C)\};$$

（3）从 $\mathscr{F}$ 中删去所有含变量 B 的函数 $\psi_1(B)$ 和 $P(C \mid B)$，并生成新函数

$$\psi_2(C) = \sum_B P(C \mid B) \psi_1(B),$$

然后将其加入 $\mathscr{F}$，得到

$$\mathscr{F} = \{\underline{\psi_2(C)}, \underline{P(D \mid C)}\};$$

（4）从 $\mathscr{F}$ 中删去所有含变量 C 的函数 $\psi_2(C)$ 和 $P(D \mid C)$，并生成新函数

$$\psi_3(D) = \sum_C P(D \mid C)\psi_2(C);$$

(5) 返回 $\psi_3(D)$，它正是 $P(D)$.

注意，在第 (2) 步以后，变量 A 不再在 $\mathscr{F}$ 的函数中出现，所以可以说，它在第 (2) 步中被消去，而 A 被消去的过程称为消元. 变量 B 和 C 分别在第 (3) 步和第 (4) 步被消去. 在消去 A，B 和 C 后，得到一个函数，它只涉及查询变量 D. 下面严格定义消元运算.

4.2.2 消元运算

第 2.2 和 2.3 两节引入了联合分布的分解的概念. 抽象地讲，一个联合分布是一个多变量函数. 于是，分解的概念可以自然地推广到一般的多元函数. 设 $F(X_1, X_2, \cdots, X_n)$ 是变量 $\{X_1, X_2, \cdots, X_n\}$ 的一个函数，而 $\mathscr{F}=\{f_1, f_2, \cdots, f_m\}$ 是一组函数，其中每个 f_i 所涉及的变量是 $\{X_1, X_2, \cdots, X_n\}$ 的一个子集. 如果

$$F = \prod_{i=1}^{m} f_i, \tag{4.3}$$

则称 $\mathscr{F}$ 是 F 的一个**分解** (factorization)，$f_1, f_2, \cdots, f_m$ 称为这个分解的**因子** (factor).

从 $F(X_1, X_2, \cdots, X_n)$ 出发，可通过如下方式获得变量 $\{X_2, \cdots, X_n\}$ 的一个函数：

$$G(X_2, \cdots, X_n) = \sum_{X_1} F(X_1, X_2, \cdots, X_n). \tag{4.4}$$

这个过程称为**消元** (elimination)：从函数 $F(X_1, X_2, \cdots, X_n)$ 中消去 X_1，得到函数 $G(X_2, \cdots, X_n)$.

设 $\mathscr{F}=\{f_1, \cdots, f_m\}$ 是函数 $F(X_1, X_2, \cdots, X_n)$ 的一个分解. 从 $\mathscr{F}$ 中消去变量 X_1 指的是如下过程：

(1) 从 $\mathscr{F}$ 中删去所有涉及 X_1 的函数 (不失一般性，设这些函数是 $\{f_1, \cdots, f_k\}$)；

(2) 将新函数 $\sum_{X_1} \prod_{i=1}^{k} f_i$ 放回 $\mathscr{F}$ 中.

定理 4.1　设 $\mathscr{F}$ 是函数 $F(X_1, X_2, \cdots, X_n)$ 的一个分解. 设 $\mathscr{F}'$ 是从 $\mathscr{F}$ 中消去 X_1 后所得的一组函数，$G(X_2, \cdots, X_n)$ 是从 $F(X_1, X_2, \cdots, X_n)$ 中消去 X_1 后所得的函数. 那么，$\mathscr{F}'$ 是 $G(X_2, \cdots, X_n)$ 的一个分解.

证明　设 $\mathscr{F}=\{f_1, \cdots, f_m\}$，而 X_1 只在 $\{f_1, f_2, \cdots, f_k\}$ 中出现. 有

$$G(X_2, \cdots, X_n) = \sum_{X_1} F(X_1, X_2, \cdots, X_n)$$

$$= \sum_{X_1} \prod_{i=1}^{m} f_i$$

$$= \sum_{X_1} \left(\prod_{i=1}^{k} f_i \prod_{i=k+1}^{m} f_i \right)$$

$$= \left(\sum_{X_1} \prod_{i=1}^{k} f_i \right) \prod_{i=k+1}^{m} f_i \quad \left(\text{因为} \prod_{i=k+1}^{m} f_i \text{ 不涉及 } X_1\right).$$

定理得证. □

很明显，直接从 $F(X_1, X_2, \cdots, X_n)$ 中消去变量 X_1 而得到 $G(X_2, \cdots, X_n)$ 的计算复杂度相对于变量个数 n 成指数增长. 但是，从 $F(X_1, X_2, \cdots, X_n)$ 的一个分解 $\mathscr{F}$ 中消去 X_1，而得到 $G(X_2, \cdots, X_n)$ 的一个分解 $\mathscr{F}'$ 的计算复杂度未必相对于 n 成指数增长. 实际上，这个过程的主要运算是计算 $\sum_{X_1} \prod_{i=1}^{k} f_i$，其复杂度只依赖于涉及 X_1 的因子 $f_1, \cdots, f_k$ 所涉及到的变量个数可能远远小于 n.

4.2.3 算法描述

本小节给出变量消元算法. 先来定义一个在函数上的运算. 设 $f(\mathbf{X}, \mathbf{Y})$ 为两组变量 $\mathbf{X}$ 和 $\mathbf{Y}$ 的函数，其中 $\mathbf{X} \cap \mathbf{Y} = \varnothing$，并设 $\mathbf{x}$ 为 $\mathbf{X}$ 的一取值. 在 $f(\mathbf{X}, \mathbf{Y})$ 中，将 $\mathbf{X}$ 设置为 $\mathbf{x}$ 得到一个关于 $\mathbf{Y}$ 的函数 $f_{\mathbf{X}=\mathbf{x}}(\mathbf{Y})$：

$$f_{\mathbf{X}=\mathbf{x}}(\mathbf{Y} = \mathbf{y}) = f(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}), \quad \forall \mathbf{y} \in \Omega_{\mathbf{Y}}. \tag{4.5}$$

为方便起见，有时把 $f_{\mathbf{X}=\mathbf{x}}(\mathbf{Y})$ 记为 $f(\mathbf{X} = \mathbf{x}, \mathbf{Y})$.

例 4.3 设 X 和 Y 是取值为 0 或 1 的二值变量，$f(X, Y)$ 是下表所给出的函数：

X \ Y	0	1
0	0.3	0.8
1	0.6	0

在 $f(X, Y)$ 中将 X 设置为 0，得到如下表所示的函数 $f(X = 0, Y)$：

Y	0	1
$f(X = 0, Y)$	0.3	0.8

□

设 $\mathbf{X}$ 是一个贝叶斯网 $\mathscr{N}$ 中所有变量的集合，$\mathscr{F}$ 是 $\mathscr{N}$ 中所有概率分布的集合. 按照贝叶斯网的定义，$\mathscr{F}$ 是 $\mathscr{N}$ 所表示的联合概率分布 $P(\mathbf{X})$ 的一个分解. 假设观测到了证据 $\mathbf{E} = \mathbf{e}$. 在 $\mathscr{F}$ 的因子中，将各证据变量设置为它们的观测值，得到另

一组函数，记之为$\mathscr{F}'$. 这一步称为**证据设置**. 不难看出，$\mathscr{F}'$是函数 $P(\mathbf{Y}, \mathbf{E}=\mathbf{e})$ 的一个分解，这里 $\mathbf{Y} = \mathbf{X} \setminus \mathbf{E}$.

设 $\mathbf{Q}$ 是 $\mathbf{Y}$ 的一个子集合. 从$\mathscr{F}'$中逐个消去所有在 $\mathbf{Y}$ 中但不在 $\mathbf{Q}$ 中的变量，得到另一个函数集合，记之为$\mathscr{F}''$. 根据定理 4.1，$\mathscr{F}''$是 $P(\mathbf{Q},\mathbf{E} = \mathbf{e})$ 的一个分解. 所以，将$\mathscr{F}''$中的所有因子相乘，就得到 $P(\mathbf{Q}, \mathbf{E} = \mathbf{e})$. 按照条件概率的定义，可以进一步得到

$$P(\mathbf{Q} \mid \mathbf{E} = \mathbf{e}) = \frac{P(\mathbf{Q}, \mathbf{E} = \mathbf{e})}{P(\mathbf{E} = \mathbf{e})}. \tag{4.6}$$

其中 $P(\mathbf{E}=\mathbf{e}) = \sum_{\mathbf{Q}} P(\mathbf{Q},\mathbf{E}=\mathbf{e})$. 上述过程实际上给出了一个计算后验概率分布 $P(\mathbf{Q} \mid \mathbf{E} = \mathbf{e})$ 的算法，即变量消元法，简称 VE 算法. 图 4.3 给出了这个算法的伪代码.

VE($\mathcal{N}$,**E**,**e**,**Q**,ρ)

输入：$\mathcal{N}$—— 一个贝叶斯网；**E** —— 证据变量；

　　e —— 证据变量的取值；**Q** —— 查询变量；

　　ρ —— 消元顺序，包含所有不在 $\mathbf{Q} \cup \mathbf{E}$ 中的变量.

输出：$P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$.

　　1：$\mathscr{F} \leftarrow \mathcal{N}$中所有概率分布的集合；

　　2：在$\mathscr{F}$的因子中，将证据变量 **E** 设置为其观测值 **e**；

　　3：**while** ($\rho \neq \phi$)

　　4：　　设 Z 为 ρ 中排在最前面的变量，将 Z 从 ρ 中删去；

　　5：　　$\mathscr{F} \leftarrow \mathrm{Elim}(\mathscr{F}, Z)$；

　　6：**end while**

　　7：将$\mathscr{F}$中所有因子相乘，得到一个 **Q** 的函数 $h(\mathbf{Q})$；

　　8：**return** $h(\mathbf{Q}) \Big/ \sum_{\mathbf{Q}} h(\mathbf{Q})$.

Elim ($\mathscr{F}$, Z)

输入：$\mathscr{F}$—— 一个函数集合；

　　Z —— 待消元变量.

输出：另一个函数集合.

　　1：从$\mathscr{F}$中删去所有涉及 Z 的函数，设这些函数是 $\{f_1, \cdots, f_k\}$；

　　2：$g \leftarrow \prod_{i=1}^{k} f_i$；

　　3：$h \leftarrow \sum_{Z} g$；

　　4：将 h 放回$\mathscr{F}$；

　　5：**return** $\mathscr{F}$.

图 4.3　变量消元算法

VE 算法有 5 个输入：①一个贝叶斯网 $\mathcal{N}$，它也是联合分布 $P(\mathbf{X})$ 的一个分解；②证据变量的集合 $\mathbf{E}$；③证据变量的取值 $\mathbf{e}$；④查询变量的集合 $\mathbf{Q}$；⑤所有不在 $\mathbf{Q} \cup \mathbf{E}$ 中的变量的一个排序 ρ，称为**消元顺序**（elimination order）. VE 算法首先设置证据（第 2 步）；然后按照顺序 ρ 逐个消去不在 $\mathbf{Q} \cup \mathbf{E}$ 中的所有变量；接着把所有 $\mathscr{F}$ 中的因子相乘，得到 $\mathbf{Q}$ 的一个函数 $h(\mathbf{Q})$；最后将 $h(\mathbf{Q})$ 规一化，得到 $P(\mathbf{Q} \mid \mathbf{E} = \mathbf{e})$. 前面两段的讨论已经证明了下面的定理.

定理 4.2 算法 VE 返回的是 $P(\mathbf{Q} \mid \mathbf{E} = \mathbf{e})$. □

4.2.4 一个例子

例 4.4 在图 4.4 所示的贝叶斯网中，设证据为 $\{F = 0\}$，考虑调用 VE 算法计算 $P(A \mid F = 0)$.

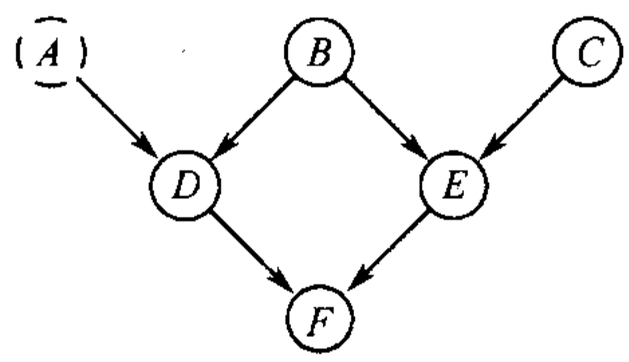


图 4.4 变量消元法示例，求 $P(A \mid F = 0)$，消元顺序为 $\langle C, E, B, D \rangle$

设变量消元顺序 $\rho = \langle C, E, B, D \rangle$. 贝叶斯网给出的联合分布的分解为

$$\mathscr{F} = \{P(A), P(B), P(C), P(D \mid A, B), P(E \mid B, C), P(F \mid D, E)\}.$$

(1) VE 算法首先设置证据 $F = 0$，得

$$\mathscr{F} = \{P(A), P(B), \underline{P(C)}, P(D \mid A, B), \underline{P(E \mid B, C)}, P(F = 0 \mid D, E)\};$$

(2) 第一个要消去的变量为 C，与之有关的函数是 $P(C)$ 和 $P(E \mid B, C)$. 消去 C，得

$$\mathscr{F} = \{P(A), P(B), P(D \mid A, B), \underline{P(F = 0 \mid D, E)}, \underline{\psi_1(B, E)}\},$$

这里 $\psi_1(B, E) = \sum_C P(C) P(E \mid B, C)$；

(3) 下一个要消去的变量为 E，与之有关的函数是 $P(F = 0 \mid D, E)$ 和 $\psi_1(B, E)$. 消去 E，得

$$\mathscr{F} = \{P(A), \underline{P(B)}, \underline{P(D \mid A, B)}, \underline{\psi_2(B, D)}\},$$

这里 $\psi'_2(B, D) = \sum_E P(F = 0 \mid D, E) \psi_1(B, E)$；

(4) 下一个要消去的变量为 B，与之有关的函数是 $P(B)$，$P(D \mid A, B)$ 和 $\psi_2(B, D)$. 消去 B，得

$$\mathscr{F} = \{P(A), \underline{\psi_3(A, D)}\},$$

这里 $\psi_3(A, D) = \sum_B P(B)\ P(D \mid A, B)\psi_2(B, D)$；

(5) 最后一个要消去的变量为 D，与之有关的函数是 $\psi_3(A, D)$. 消去 D，得

$$\mathscr{F} = \{P(A), \psi_4(A)\},$$

这里 $\psi_4(A) = \sum_D \psi_3(A, D)$；

(6) 计算 $h(A) = P(A)\ \psi_4(A)$；

(7) 返回

$$P(A \mid F = 0) = \frac{h(A)}{\sum_A h(A)}.$$

这即是要求的后验概率分布. □

4.3　复杂度分析

本节分析 VE 算法的复杂度. 我们首先讨论如何度量复杂性，然后给出一个计算复杂度的方法.

4.3.1　复杂性的度量

在 VE 算法中，最耗费时间和空间的步骤是对 Elim($\mathscr{F}$, Z) 的调用. 它的运算复杂度远远超过其它步骤，所以可以把这一步的复杂度作为整个算法的复杂度.

Elim($\mathscr{F}$, Z) 从 $\mathscr{F}$ 中挑出所有涉及 Z 的函数，$\{f_1, \cdots, f_k\}$，将它们相乘得到中间函数 g，再将 Z 从 g 中消去. 设 $X_1, \cdots, X_l$ 是 g 中所有除 Z 之外的变量. 如果把函数表示成多维表，则 g 所需储存的函数值的个数为 $|Z| \prod_{i=1}^{l} |X_i|$. 因此，$|Z| \prod_{i=1}^{l} |X_i|$ 是函数 g 的复杂性的一个恰当的度量，从而也是 Elim($\mathscr{F}$, Z) 的复杂性的一个恰当的度量. 我们称之为变量 Z 的**消元成本** (elimination cost)，记为 $\mathscr{C}(Z)$.

在推理过程中，VE 算法需要消去多个变量，设它们依次为 $Z_1, Z_2, \cdots, Z_n$. 我们用总消元成本 $\mathscr{C} = \sum_{i=1}^{n} \mathscr{C}(Z_i)$ 来度量 VE 算法的复杂度. 式 $\sum_{i=1}^{n} \mathscr{C}(Z_i)$ 中最大的一项对整个式子的值的大小起支配作用，因此有时也用它来度量 VE 算法的复杂性①.

例 4.5　在例 4.4 中，设网中所有变量均取二值，消元顺序为 $\rho = \langle C, E, B, D \rangle$. 在消去 C 时，需要计算 $g = P(C)\ P(E \mid B, C)$，共涉及 3 个变量：B, C 和

① 这个最大项往往被称为**最大团大小** (maximum clique size)，在第 5 章中将会看到此名称的由来.

E. 所以 C 的消元成本为 $\mathscr{C}(C)=2\times2\times2=8$. 类似地，其它 3 个节点 E, B 和 D 的消元成本分别为 8，8 和 4. 因此，用 ρ 进行变量消元的总成本为 28，其中最大一项为 8. □

4.3.2 复杂度的计算

下面说明怎样通过分析贝叶斯网的网络结构计算 VE 算法的复杂度.

1. 结构图与端正图

考虑一个函数的集合 $\mathscr{F}=\{f_1,\cdots,f_m\}$，$\mathscr{F}$ 的**结构图**（structural graph）是一个按如下方法定义的无向图：

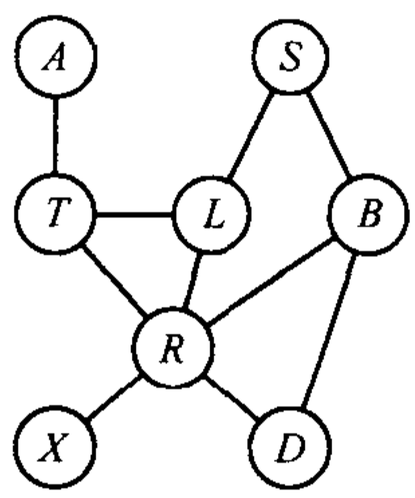
(1) 从空图出发，对 $\mathscr{F}$ 中的每一个变量，在图中相应地添加一个节点；

(2) 对任意两个变量 X 和 Y，如果它们出现在同一个因子 f_i 中，则在它们对应的节点之间添加一条边.

例如，函数集合

$$\mathscr{F}=\{P(A),\ P(T\mid A),\ P(S),\ P(L\mid S),P(B\mid S),\\ P(R\mid T,L),\ P(X\mid R),\ P(D\mid R,B)\}$$

的结构图如图 4.5 所示. 其中因为 T 和 A 都出现在 $P(T\mid A)$ 中，所以有边“T-A”；其余各边以此类推.



$\mathscr{F}=\{P(A),P(T|A),P(S),P(L|S),P(B|S),P(R|T,L),P(X|R),P(D|R,B)\}$

图 4.5 函数集合 $\mathscr{F}$ 的结构图

注意，上述函数集合 $\mathscr{F}$ 是由图 2.5 (b) 所示的贝叶斯网 $\mathscr{N}$ 中的概率分布所组成的，而它的结构图正好是 $\mathscr{N}$ 的端正图. 实际上，任何一个贝叶斯网中所有概率分布函数的集合的结构图与该网络的端正图相同.

2. 结构图与消元复杂度

从一组函数 $\mathscr{F}$ 中消去一个变量 Z 的成本可以从 $\mathscr{F}$ 的结构图中算出. 如前所

述，从 $\mathscr{F}$ 中消去 Z 的成本是 $\mathscr{C}(Z)=|Z|\prod_{i=1}^{l}|X_i|$，其中 $X_1,\cdots,X_l$，Z 是函数 $g=\prod_{i=1}^{k}f_i$ 所涉及到的所有变量，而 $f_1,\cdots,f_k$ 是所有涉及 Z 的因子. 根据结构图的定义，$\{X_1,\cdots,X_l\}$ 刚好是在 $\mathscr{F}$ 的结构图中所有与 Z 相邻的节点的集合，记为 nb(Z). 因此，有

$$\mathscr{C}(Z)=|Z|\prod_{X\in \mathrm{nb}(Z)}|X|. \tag{4.7}$$

例 4.6　考虑图 4.5 中的 $\mathscr{F}$ 及其结构图. 设所有变量均取二值. 从 $\mathscr{F}$ 中消去变量 T 需要计算 $g=P(T\mid A)P(R\mid T,L)$，涉及到 4 个节点 $\{T,A,R,L\}$，所以

$$\mathscr{C}(T)=|T|\prod_{i=1}^{l}|X_i|=|T|\cdot|A|\cdot|R|\cdot|L|=2\times2\times2\times2=16.$$

而在其结构图中，与 T 相邻的节点是 $\mathrm{nb}(T)=\{A,R,L\}$. 所以 $\mathscr{C}(T)$ 也可以按式 (4.7) 从结构图中计算出来，即

$$\mathscr{C}(T)=|T|\prod_{X\in \mathrm{nb}(T)}|X|=16\ .$$

□

3. 消元与结构图变换

设 $\mathscr{G}$ 为 $\mathscr{F}$ 的结构图，$\mathscr{F}'$ 为从 $\mathscr{F}$ 中消去变量 Z 后得到的函数集合. $\mathscr{F}'$ 的结构图可以通过在 $\mathscr{G}$ 上做如图 4.6 变换而得到：

(1) 在 $\mathscr{G}$ 中将所有与 Z 相邻的节点 nb(Z) 两两相连；

(2) 从图中除去节点 Z.

上述过程称为从 $\mathscr{G}$ 中消去 Z 的图消元过程，记为 Elim($\mathscr{G}$, Z).

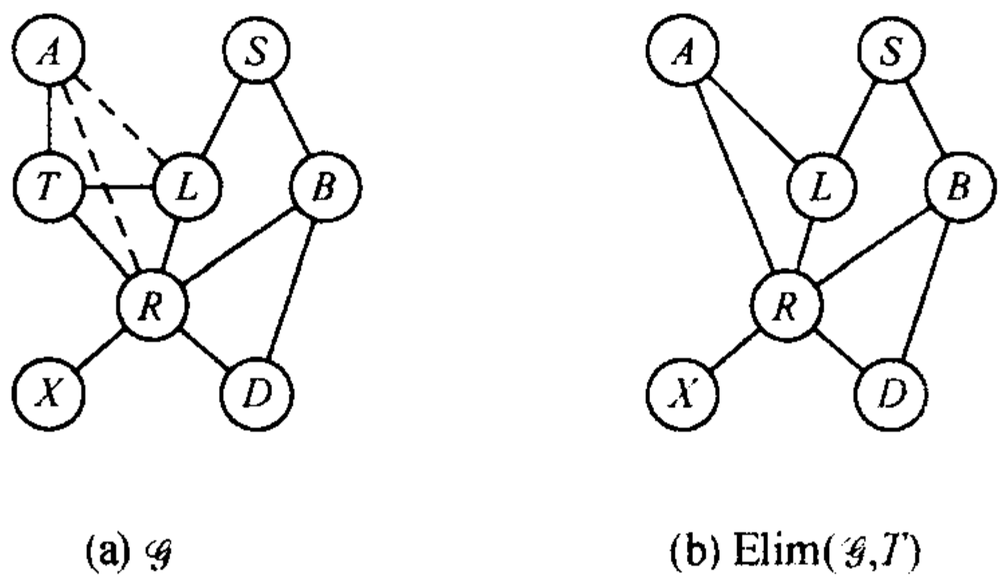
例 4.7　考虑图 4.5 中的函数集合 $\mathscr{F}$ 所对应的结构图 $\mathscr{G}$. 我们从中消去变量 T：先在 T 周围添加边 "A-L" 和 "A-R"，然后从图中将 T 除去. 此过程及其结果分别如图 4.6(a)和 (b) 所示. 另一方面，从 $\mathscr{F}$ 中消去 T 所得到的分解为

$$\mathscr{F}'=\{P(A),\ P(S),\ P(L\mid S),P(B\mid S),\ P(X\mid R),\ P(D\mid R,\ B),\ \psi(A,\ L,\ R)\},$$

其中 $\psi(A,\ L,\ R)=\sum_T P(T\mid A)P(R\mid T,\ L)$. 不难看出，$\mathscr{F}'$ 的结构图正好是 Elim($\mathscr{G}$, T) 产生的无向图.　□

4. VE 算法的复杂度

考虑用 VE 算法在一个贝叶斯网 $\mathscr{N}$ 中按消元顺序 ρ 计算后验概率 $P(\mathbf{Q}\mid\mathbf{E}=\mathbf{e})$. 它的复杂度是多少？图 4.7 所示的 CostVE 算法回答了这个问题. 它首先构



(a) $\mathscr{G}$ (b) Elim($\mathscr{G}$,T)

图 4.6 图消元 Elim ($\mathscr{G}$, T)

造 $\mathscr{N}$的 端正图 $\mathscr{G}$. 根据前面的讨论，$\mathscr{G}$ 正好是 VE 算法第 1 行所定义的函数集合 $\mathscr{F}$的结构图. 而CostVE算法的第 2 行和 VE 算法的第 2 行正好对应，之后 $\mathscr{G}$ 仍然是此时 $\mathscr{F}$的结构图. 接下来消去 ρ 中的第 1 个变量 Z. 根据前面的讨论，CostVE 算法的第 6 行所给出的量正好是从 $\mathscr{F}$ 中消去 Z 的成本. 而且，Elim($\mathscr{G}$, Z) 又是 Elim($\mathscr{F}$, Z) 的结构图，所以，CostVE 算法的 while 循环所计算的正好是 VE 算法的 while 循环的复杂度. 按前面的定义，这正是 VE 算法的复杂度.

CostVE($\mathscr{N}$,**E**,**Q**,ρ)

输入:$\mathscr{N}$ —— 贝叶斯网结构;**E** —— 证据变量;

Q —— 查询变量;ρ —— 消元顺序,包含所有不在 **Q**∪**E** 中的变量.

输出:用 VE 算法计算 $P(\mathbf{Q}|\mathbf{E}=\mathbf{e})$的复杂度.

1:$\mathscr{G} \leftarrow \mathscr{N}$的端正图;

2:从 $\mathscr{G}$ 中除去所有证据变量;

3:$\mathscr{C} \leftarrow 0$;

4:**while**($\rho \neq \varnothing$)

5: 从 ρ 中移去第 1 个变量 Z;

6: $\mathscr{C} \leftarrow \mathscr{C} + |Z| \prod_{X \in \mathrm{nb}(Z)} |X|$;

7: $\mathscr{G} \leftarrow \mathrm{Elim}(\mathscr{G}, Z)$;

8:**end while**

9:**return** $\mathscr{C}$.

图 4.7 算法 CostVE：从网络结构计算 VE 算法的复杂度

例 4.8 考虑用 VE 算法消去图 2.5(b)所示贝叶斯网中的所有变量. 设这些变量均取二值. 我们用 CostVE 算法来计算两个不同的消元顺序分别对应的消元成本.

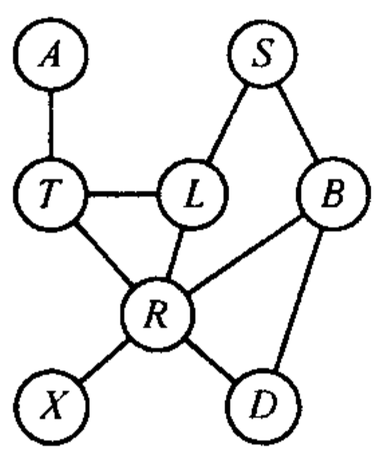
(1) $\rho_1 = \langle A, X, D, S, B, L, T, R \rangle$：CostVE 算法对应的图变换过程如图 4.8 所示．首先构造网络结构的端正图，结果如图 4.8(a)所示．然后从图中依次消去 ρ_1 中各变量，消元成本分别是 $2^2, 2^2, 2^3, 2^3, 2^3, 2^3, 2^2, 2$，分别如图 4.8(a)～(h)所示．所以，$\rho_1$ 的消元总成本为

$$\mathscr{C} = 2^2 + 2^2 + 2^3 + 2^3 + 2^3 + 2^3 + 2^2 + 2 = 46.$$

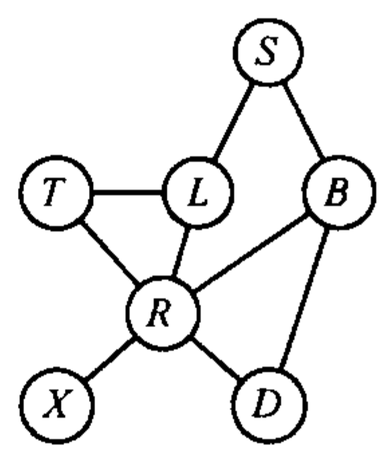
(2) $\rho_2 = \langle R, L, T, D, B, S, A, X \rangle$：其图变换过程如图 4.9 所示，消元总成本为

$$\mathscr{C} = 2^6 + 2^6 + 2^6 + 2^5 + 2^4 + 2^3 + 2^2 + 2 = 254.$$

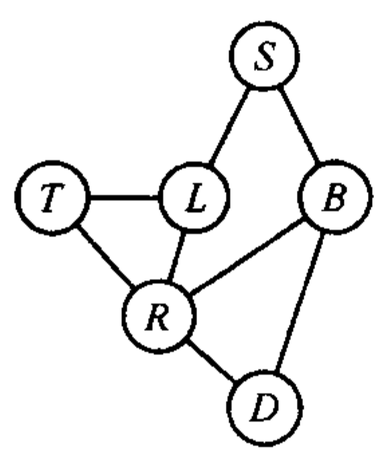
□



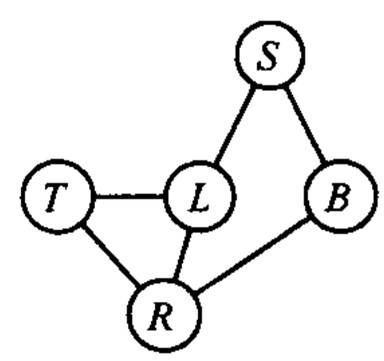
(a) 消去A，成本$\mathscr{C}(A)=2^2$



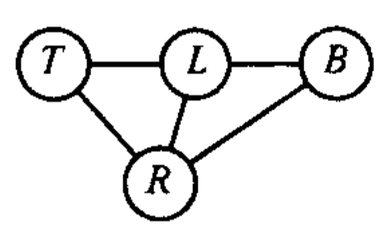
(b) 消去X，成本$\mathscr{C}(X)=2^2$



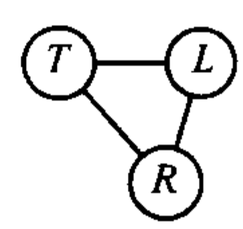
(c) 消去D，成本$\mathscr{C}(D)=2^3$



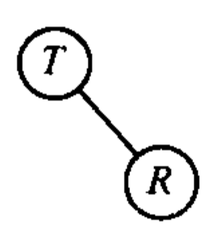
(d) 消去S，成本$\mathscr{C}(S)=2^3$



(e) 消去B，成本$\mathscr{C}(B)=2^3$



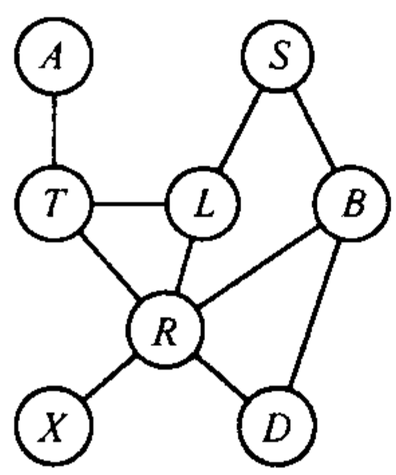
(f) 消去L，成本$\mathscr{C}(L)=2^3$



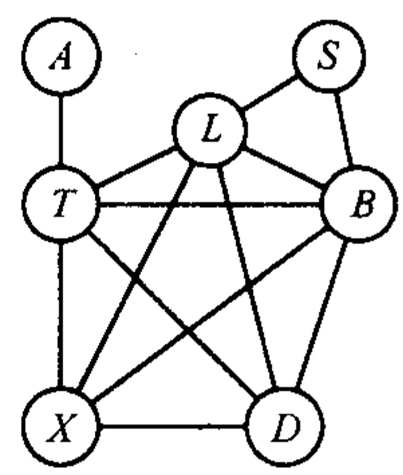
(g) 消去T，成本$\mathscr{C}(T)=2^2$

(h) 消去R，成本$\mathscr{C}(R)=2$

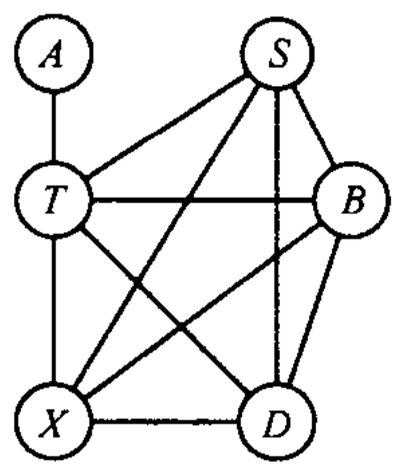
图 4.8　消元顺序 $\rho_1 = \langle A, X, D, S, B, L, T, R \rangle$ 的总成本为 46



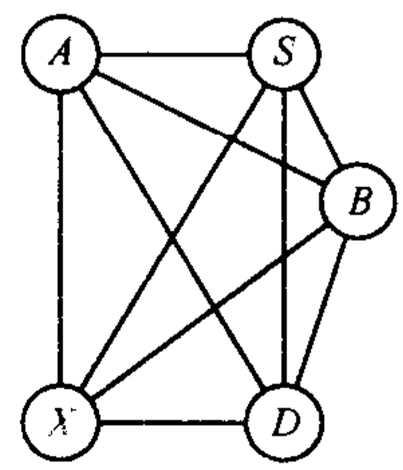
(a) 消去R, 成本$\mathscr{C}(R)=2^6$



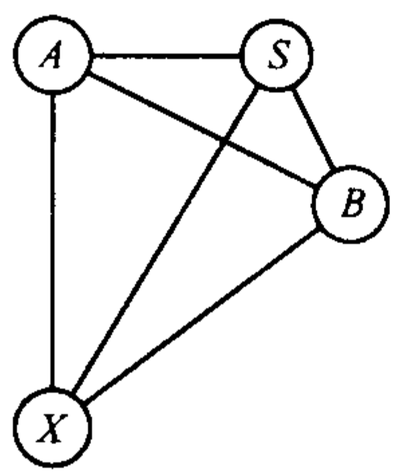
(b) 消去L, 成本$\mathscr{C}(L)=2^6$



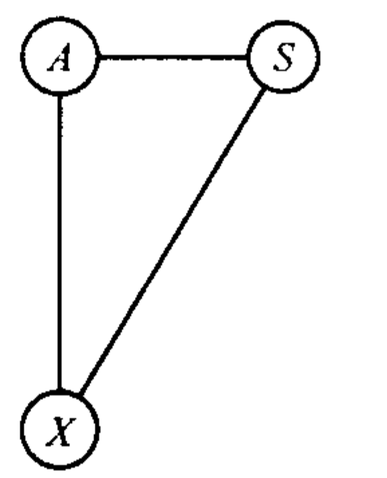
(c) 消去T, 成本$\mathscr{C}(T)=2^6$



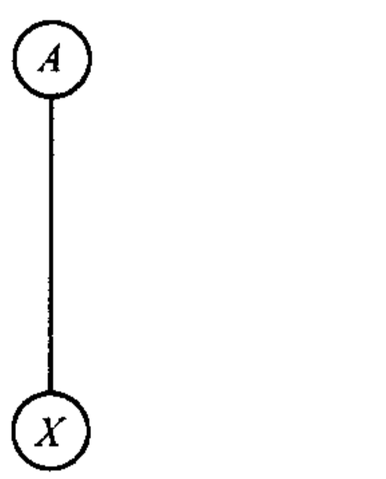
(d) 消去D, 成本$\mathscr{C}(D)=2^5$



(e) 消去B, 成本$\mathscr{C}(B)=2^4$



(f) 消去S, 成本$\mathscr{C}(S)=2^3$



(g) 消去A, 成本$\mathscr{C}(A)=2^2$

(h) 消去X, 成本$\mathscr{C}(X)=2$

图 4.9　消元顺序 $\rho_2 = \langle R, L, T, D, B, S, A, X \rangle$ 的总成本为 254

4.4　消元顺序

从例 4.8 可见，不同的消元顺序可能会导致不同的计算复杂度．总成本最低的消元顺序称为**最优消元顺序**（optimal elimination order）．很自然，我们希望

能在推理之前预先知道最优消元顺序. 然而，寻找最优消元顺序是一个 NP-难解问题. 实际中，人们使用启发式规则来寻找较好、但却不一定是最优的消元顺序. 本节介绍两个常用的启发式算法，即**最大势搜索**（maximum cardinality search）和**最小缺边搜索**（minimum deficiency search）.

4.4.1　最大势搜索

设 $\mathscr{G}$ 是一个包含 n 个节点的无向图. 最大势搜索对 $\mathscr{G}$ 中所有节点按如下规则从大到小进行编号：在第 i 步中，选择拥有最多已编号相邻节点的未编号节点，将其编号为 $n-i+1$. 若这样的节点有多个，则任选其一. 在所有的节点均被编号以后，按编号由小到大将节点排序，就得到一个变量消元顺序.

例 4.9　图 4.10 所示的无向图共有 8 个节点，用最大势搜索将它们排序. 第 1 步任选一个节点，例如 B，把它编为 8 号. 第 2 步可选的节点是 S，D 和 R，因为只有它们有 1 个已编号相邻节点，即 B. 从中任选一个，设其为 S，把它编为 7 号. 第 3 步可选的节点包括 D，R 和 L，它们的已编号相邻节点的个数都是 1. 任选一个，设其为 D，把它编为 6 号. 第 4 步只能选 R，因为它的已编号邻居最多，有两个，分别是 B 和 D. 我们把 R 编为 5 号. 接着，因为同样的理由，第 5 步只能选 L，把它编为 4 号. 按照同样的方法，剩余的节点 T，X 和 A 分别被编为 3，2 和 1 号. 最后得到的消元顺序为 $\langle A, X, T, L, R, D, S, B\rangle$. 读者可以用 CostVE 算法计算一下这个顺序的消元成本. □

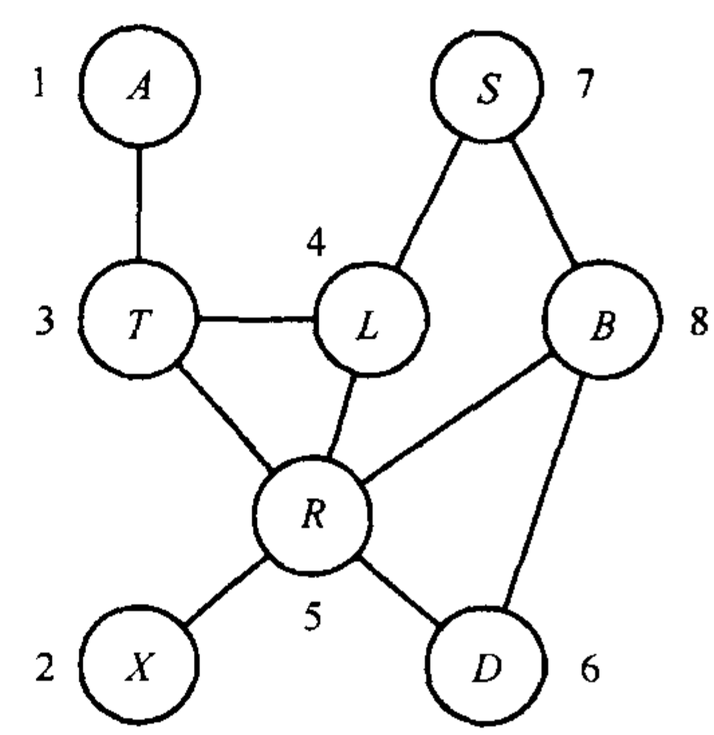


图 4.10　最大势搜索示例：所得消元顺序为 $\langle A,X,T,L,R,D,S,B\rangle$

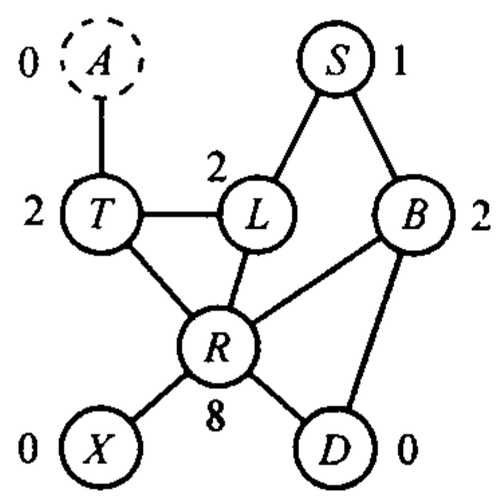
4.4.2　最小缺边搜索

一个无向图 $\mathscr{G}$ 中的某个节点 Z 的**缺边数**（deficiency）是在用 Elim($\mathscr{G}$, Z) 消去 Z 时需要添加的边的条数. 比如，在图 4.11(a) 所示的网络中，各节点 $\{A, T, L, R, S, B, D, X\}$ 的缺边数分别为 $\{0, 2, 2, 8, 1, 2, 0, 0\}$.

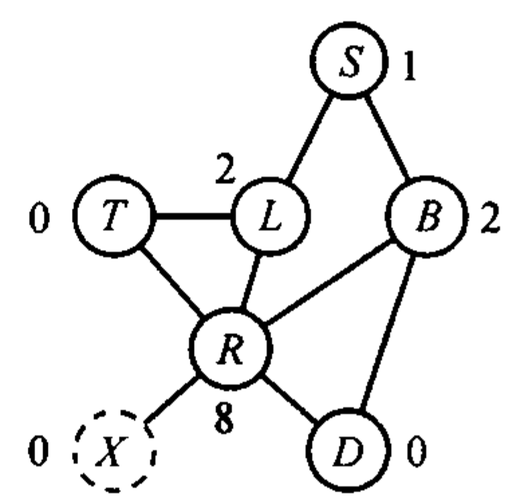
最小缺边搜索一边为 $\mathscr{G}$ 中的节点排序，一边把它们逐个从 $\mathscr{G}$ 中消去. 它在每一步都计算所有节点的缺边数，选择缺边数最小的那个节点作为下一个消元节点，同时将其从图 $\mathscr{G}$ 中消去. 当有多个节点的缺边数相等时，任选其一. 如此循环，直到所有节点都被消去. 图 4.11 展示的是最小缺边搜索在图 4.11(a)所示

的无向图上的运行情况，所获得的序为〈A, X, D, T, S, L, B, R〉.

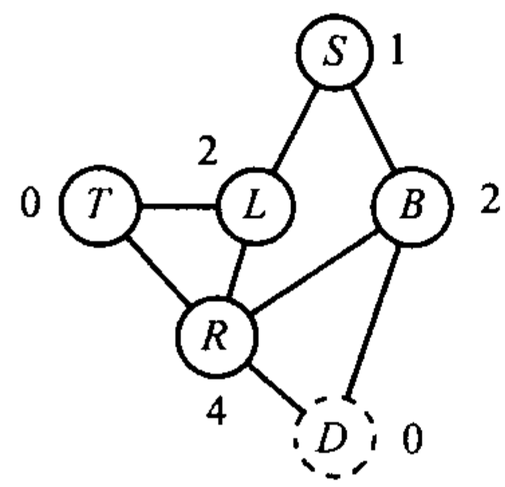
经验表明，最小缺边搜索所给出的消元顺序往往优于其它方法所给出的顺序.



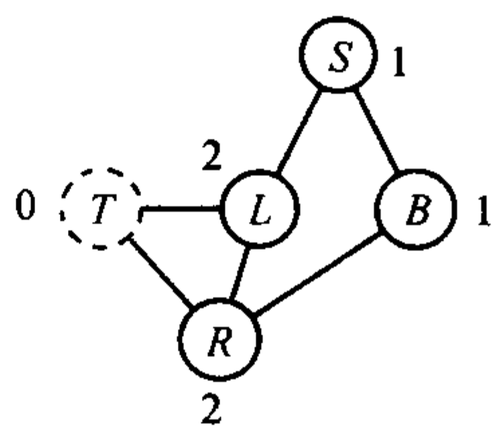
(a) A,X和D的缺边数都是0,同为最小.
任选其一，如A，将其从图中消去



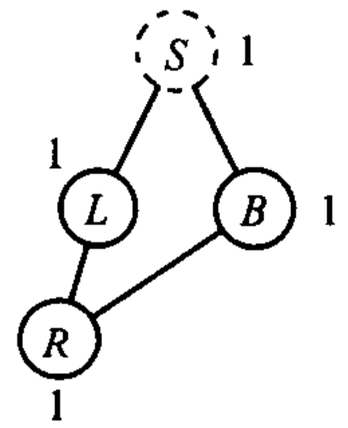
(b) X,T和D的缺边数都是0,同为最小.
任选其一，如X，将其从图中消去



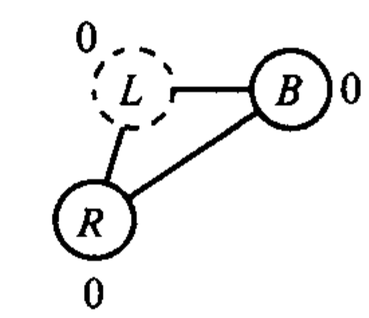
(c) T和D的缺边数都是0,同为最小.
任选其一，如D，将其从图中消去



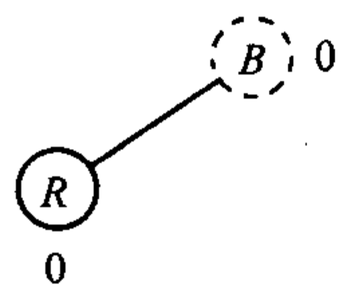
(d) T的缺边数是0,为最小.
将其从图中消去



(e) 所有节点的缺边数都相同. 任选其一，如S，将其从图中消去



(f) 所有节点的缺边数都相同. 任选其一，如L，将其从图中消去



(g) 所有节点的缺边数都相同. 任选其一，如B，将其从图中消去

(h) 唯一节点R，排第8

图 4.11　最小缺边搜索：所得消元顺序为〈A,X,D,T,S,L,B,R〉

4.5　推理问题简化

通过对网络结构的分析，往往可以找出一些与推理无关的变量. 在进行推理之前，从网络中先除去这样的变量，这样不仅不会影响推理的结果，而且能够简化推理问题，缩短推理时间.

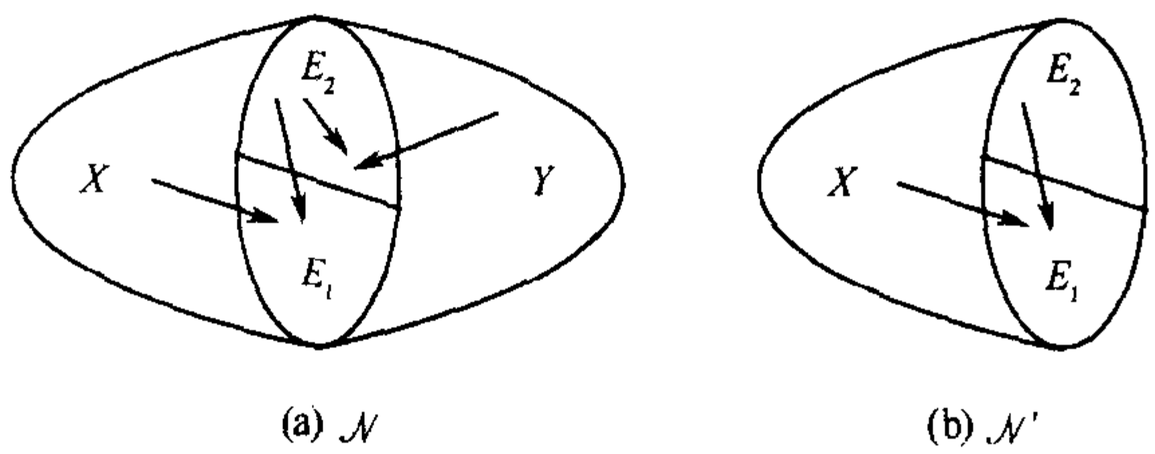


图 4.12　利用条件独立简化贝叶斯网推理

定理 4.3　给定一贝叶斯网 $\mathcal{N}$, 设 $P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$ 是需要计算的后验概率. 令 $\mathcal{N}'$为从 $\mathcal{N}$中除去在 $\mathrm{an}(\mathbf{Q} \cup \mathbf{E})$ 之外的所有节点后所得的贝叶斯网，则

$$P_{\mathcal{N}}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}) = P_{\mathcal{N}'}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}) .$$

证明　由命题 3.1 可知

$$P_{\mathcal{N}}(\mathrm{an}(\mathbf{Q} \cup \mathbf{E})) = P_{\mathcal{N}'}(\mathrm{an}(\mathbf{Q} \cup \mathbf{E})),$$

因此有

$$P_{\mathcal{N}}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}) = P_{\mathcal{N}'}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}).$$

命题得证. □

上面的定理说的是，在 $\mathrm{an}(\mathbf{Q} \cup \mathbf{E})$ 之外的所有节点与 $P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$ 无关，所以可以在推理前先将它们除去. 下面的定理能让我们除去更多的与 $P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$ 无关的节点. 在此之前，需要先定义 m-**分隔**的概念.

设 $\mathbf{X}$，$\mathbf{Y}$ 和 $\mathbf{Z}$ 为有向无圈图 $\mathcal{G}$ 中的 3 个两两交空的节点集合. 如果 $\mathbf{Z}$ 在 $\mathcal{G}$ 的端正图中 u-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$，则称 $\mathbf{Z}$ 在 $\mathcal{G}$ 中 m-分隔 $\mathbf{X}$ 和 $\mathbf{Y}$.

定理 4.4　给定一贝叶斯网 $\mathcal{N}$, 设 $P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$ 是需要计算的后验概率. 设 $\mathbf{Y}$ 是被 $\mathbf{E}$ 将其与 $\mathbf{Q}$ m-分隔的所有节点的集合，$\mathbf{X}$ 为除 $\mathbf{E}$ 和 $\mathbf{Y}$ 之外的所有节点的集合，$\mathbf{E}_1$ 为 $\mathbf{E}$ 中所有父节点在 $\mathbf{X}$ 中的节点的集合，$\mathbf{E}_2 = \mathbf{E} \setminus \mathbf{E}_1$. 按如下规则构造新的贝叶斯网$\mathcal{N}'$:

(1) 从 $\mathcal{N}$中除去所有 $\mathbf{Y}$ 中节点；

(2) 对 $\mathbf{E}_2$ 中所有非根节点，除去所有到它的边，然后将其概率分布设为均匀分布.

则有

$$P_{\mathcal{N}}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}) = P_{\mathcal{N}'}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}).$$

证明 由 $\mathbf{X}$ 的定义可知 $\mathbf{Q} \subseteq \mathbf{X}$. 因此只需证明

$$P_{\mathcal{N}}(\mathbf{X} \mid \mathbf{E}) = P_{\mathcal{N}'}(\mathbf{X} \mid \mathbf{E}).$$

由 $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{E}_1$ 和 $\mathbf{E}_2$ 的定义还可以看出，$\pi(\mathbf{X} \cup \mathbf{E}_1) \subseteq \mathbf{X} \cup \mathbf{E}$, $\pi(\mathbf{Y} \cup \mathbf{E}_2) \subseteq \mathbf{Y} \cup \mathbf{E}$. 因此，

$$\begin{aligned}
P_{\mathcal{N}}(\mathbf{X} \mid \mathbf{E}) &= \frac{\sum_{\mathbf{Y}} P_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}, \mathbf{E})}{\sum_{\mathbf{X},\mathbf{Y}} P_{\mathcal{N}}(\mathbf{X}, \mathbf{Y}, \mathbf{E})} \\
&= \frac{\sum_{\mathbf{Y}} f(\mathbf{X},\mathbf{E}) g(\mathbf{Y},\mathbf{E})}{\sum_{\mathbf{X},\mathbf{Y}} f(\mathbf{X},\mathbf{E}) g(\mathbf{Y},\mathbf{E})} \\
&= \frac{f(\mathbf{X},\mathbf{E}) \sum_{\mathbf{Y}} g(\mathbf{Y},\mathbf{E})}{\left[\sum_{\mathbf{X}} f(\mathbf{X},\mathbf{E})\right]\left[\sum_{\mathbf{Y}} g(\mathbf{Y},\mathbf{E})\right]} \\
&= \frac{f(\mathbf{X},\mathbf{E})}{\sum_{\mathbf{X}} f(\mathbf{Y},\mathbf{E})};
\end{aligned}$$

其中，

$$f(\mathbf{X},\mathbf{E}) = \prod_{W \in \mathbf{X} \cup \mathbf{E}_1} P_{\mathcal{N}}(W \mid \pi(W)),$$

$$g(\mathbf{Y},\mathbf{E}) = \prod_{W \in \mathbf{Y} \cup \mathbf{E}_2} P_{\mathcal{N}}(W \mid \pi(W)).$$

另一方面，

$$\begin{aligned}
P_{\mathcal{N}'}(\mathbf{X} \mid \mathbf{E}) &= \frac{P_{\mathcal{N}'}(\mathbf{X},\mathbf{E})}{\sum_{\mathbf{X}} P_{\mathcal{N}'}(\mathbf{X},\mathbf{E})} \\
&= \frac{\prod_{W \in \mathbf{X} \cup \mathbf{E}_1} P_{\mathcal{N}'}(W \mid \pi(W)) \prod_{W \in \mathbf{E}_2} P_{\mathcal{N}'}(W)}{\sum_{\mathbf{X}} \prod_{W \in \mathbf{X} \cup \mathbf{E}_1} P_{\mathcal{N}'}(W \mid \pi(W)) \prod_{W \in \mathbf{E}_2} P_{\mathcal{N}'}(W)} \\
&= \frac{\prod_{W \in \mathbf{X} \cup \mathbf{E}_1} P_{\mathcal{N}'}(W \mid \pi(W))}{\sum_{\mathbf{X}} \prod_{W \in \mathbf{X} \cup \mathbf{E}_1} P_{\mathcal{N}'}(W \mid \pi(W))}.
\end{aligned}$$

由 $\mathcal{N}'$ 的构造方法可知，对任意 $W \in \mathbf{X} \cup \mathbf{E}_1$，有 $P_{\mathcal{N}}(W \mid \pi(W)) = P_{\mathcal{N}'}(\mathbf{W} \mid \pi(\mathbf{W}))$. 故 $P_{\mathcal{N}}(\mathbf{X} \mid \mathbf{E}) = P_{\mathcal{N}'}(\mathbf{X} \mid \mathbf{E})$. 命题得证. □

例 4.10 在如图 4.13(a)所示的贝叶斯网 $\mathcal{N}_1$，设证据为 $\{B=0, F=1\}$，考

虑计算 $P_{\mathcal{N}_1}(I \mid B=0, F=1)$. 在 $\mathcal{N}_1$ 中，$\mathrm{an}(\{I, B, F\}) = \{A,B,C,D,E,F,I\}$. 应用定理 4.3，可以除去 $\mathrm{an}(I, B, F)$ 之外的节点，从而得到如图 4.13(b)所示的网络 $\mathcal{N}_2$，且有 $P_{\mathcal{N}_2}(I \mid B=0, F=1) = P_{\mathcal{N}_1}(I \mid B=0, F=1)$. 在 $\mathcal{N}_2$ 中，根据定理 4.4，又可以除去节点 A，D，最终得到如图 4.13(c)所示的网络 $\mathcal{N}_3$，且有 $P_{\mathcal{N}_3}(I \mid B=0, F=1) = P_{\mathcal{N}_2}(I \mid B=0, F=1)$. □

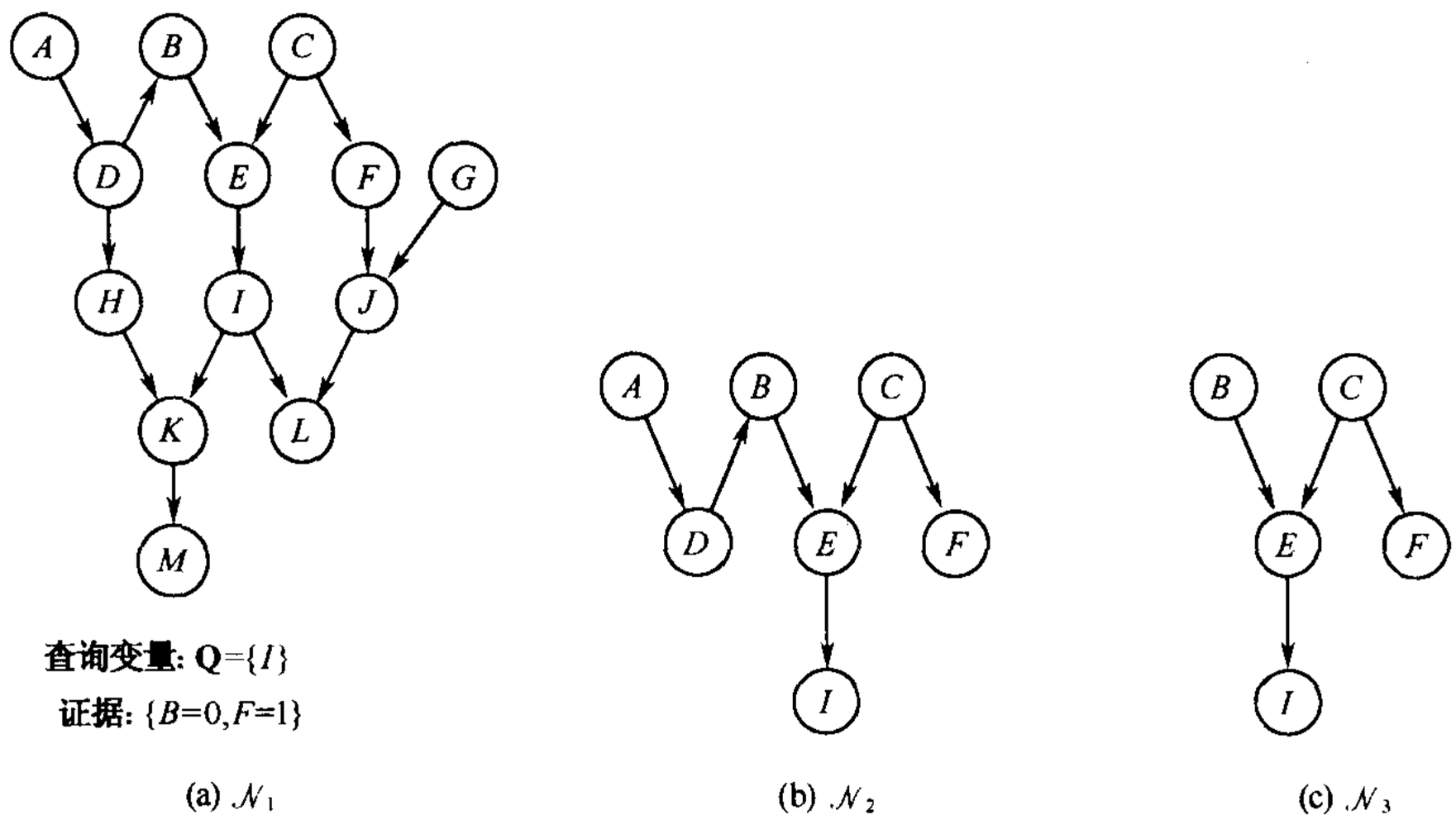


图 4.13　推理问题简化示例

4.6　MAP 假设问题

这一节考虑计算 $\arg\max_{\mathbf{h}} P(\mathbf{H}=\mathbf{h} \mid \mathbf{E}=\mathbf{e})$ 的问题. 为了解决这个问题，可以首先从贝叶斯网所提供的联合分布的分解出发，先设置证据 $\mathbf{E}=\mathbf{e}$，再利用消元法逐个消去假设变量 $\mathbf{H}$ 之外的变量，而获得一个 $P(\mathbf{H},\mathbf{E}=\mathbf{e})$ 的分解. 之后如果假设变量很少，可以把分解中的所有因子相乘得到 $P(\mathbf{H},\mathbf{E}=\mathbf{e})$ 的一个显式表示，然后利用

$$\begin{aligned}\arg\max_{\mathbf{h}} P(\mathbf{H} \mid \mathbf{E}=\mathbf{e}) &= \arg\max_{\mathbf{h}} \frac{P(\mathbf{H},\mathbf{E}=\mathbf{e})}{\sum_{\mathbf{h}} P(\mathbf{H},\mathbf{E}=\mathbf{e})} \\ &= \arg\max_{\mathbf{h}} P(\mathbf{H}, \mathbf{E}=\mathbf{e}) \qquad (4.8)\end{aligned}$$

得到结果.

上述方法的复杂度是假设变量个数 $|\mathbf{H}|$ 的指数函数，当假设变量较多时不可行. 因此，需要考虑利用 $P(\mathbf{H},\mathbf{E}=\mathbf{e})$ 的分解来降低计算复杂度. 这是如下的一个**不定序动态规划**（nonserial dynamic programming）优化问题，简称 **NDP 问题**的一个特例.

给定：函数 $f(\mathbf{H})$ 的一个分解 $\mathscr{F}$，求：

- 函数 f 的最大值 $\max\limits_{\mathbf{H}} f(\mathbf{H})$；
- 函数 f 的最大值点 $\arg\max\limits_{\mathbf{H}} f(\mathbf{H})$.

因此，本节以下的讨论将以这个问题为对象.

4.6.1 两个运算

怎样利用分解来降低 NDP 问题的计算复杂度？在回答这一问题之前，先来考虑两个运算，即最大值运算 max 和最大值点运算 arg max①.

考虑图 4.14 (a) 所示的函数 $g(X,Y)$. 如果对它的两个变量 X 和 Y 进行最大值运算，得到的是它的最大值 0.6，即

$$\max_{X,Y} g(X, Y) = 0.6.$$

如果只对其中的一个变量，比方说 X，进行最大值运算，则得到的是一个如图 4.14(b)所示的 Y 的函数，即

$$g'(Y) = \max_{X} g(X, Y).$$

其中 $g'(Y=0)$ 是函数 $g(X,Y=0)$ 的最大值，即 0.6；而 $g'(Y=1)$ 是函数 $g(X,Y=1)$ 的最大值，即 0.5. 类似地，$\max\limits_{Y} g(X, Y)$ 是变量 X 的一个函数，如图 4.14(c)所示.

Y \ X	0	1
0	0.3	0.6
1	0.4	0.5

(a) $g(X,Y)$

Y	$\max\limits_{X} g(X,Y)$
0	0.6
1	0.5

(b)

X	0	1
$\max\limits_{Y} g(X,Y)$	0.4	0.6

(c)

Y	$\arg\max\limits_{X} g(X,Y)$
0	1
1	1

(d)

X	0	1
$\arg\max\limits_{Y} g(X,Y)$	1	0

(e)

图 4.14 最大值运算 max 和最大值点运算 arg max

① 为方便计，这里假设 arg max 运算取单值.

如果对函数 $g(X,Y)$ 的两个变量 X 和 Y 进行最大值点运算，得到的是使 g 达到最大值的 X 和 Y 的取值，即

$$\arg\max_{X,Y} g(X,Y) = (1,\ 0).$$

如果只对其中的一个变量，比方说 X，进行最大值点运算，得到的是一个从 Y 的值域 Ω_Y 到 X 的值域 Ω_X 的映射，即

$$\delta(Y) = \arg\max_{X} g(X,Y).$$

其中 $\delta(Y=0)$ 是使函数 $g(X,Y=0)$ 达到最大的 X 的取值，即 1；而 $\delta(Y=1)$ 是使函数 $g(X,Y=1)$ 达到最大的 X 的取值，也是 1，如图 4.14(d)所示. 同样，$\arg\max_{Y} g(X,Y)$是一个从 X 的值域 Ω_X 到 Y 的值域 Ω_Y 的映射，如图 4.14(e)所示.

一般地，设 $\mathbf{X}$ 和 $\mathbf{Y}$ 是两个交空的变量集合，$g(\mathbf{X},\mathbf{Y})$ 是它们的函数，那么，

(1) $\max_{\mathbf{X}} g(\mathbf{X},\mathbf{Y})$ 是 $\mathbf{Y}$ 的函数；

(2) $\arg\max_{\mathbf{X}} g(\mathbf{X},\mathbf{Y})$ 是从 $\Omega_{\mathbf{Y}}$ 到 $\Omega_{\mathbf{X}}$ 的映射.

定理 4.5　设 $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ 是 3 个两两交空的变量集合，$g(\mathbf{X}, \mathbf{Y})$ 和 $h(\mathbf{Y}, \mathbf{Z})$ 是两个函数，则有

$$\max_{\mathbf{X},\mathbf{Y},\mathbf{Z}} g(\mathbf{X}, \mathbf{Y})h(\mathbf{Y},\mathbf{Z}) = \max_{\mathbf{X},\mathbf{Y}}[g(\mathbf{X}, \mathbf{Y}) \max_{\mathbf{Z}} h(\mathbf{Y}, \mathbf{Z})]. \tag{4.9}$$

再设

$$\delta(\mathbf{Y}) = \arg\max_{\mathbf{Z}} h(\mathbf{Y},\mathbf{Z}),$$

和

$$(\mathbf{x}^*, \mathbf{y}^*) = \arg\max_{\mathbf{X},\mathbf{Y}}[g(\mathbf{X},\mathbf{Y}) \max_{\mathbf{Z}} h(\mathbf{Y},\mathbf{Z})],$$

那么，

$$\arg\max_{\mathbf{X},\mathbf{Y},\mathbf{Z}} g(\mathbf{X},\mathbf{Y})\ h(\mathbf{Y},\mathbf{Z}) = (\mathbf{x}^*, \mathbf{y}^*, \delta(\mathbf{y}^*)). \tag{4.10}$$

证明　设 $F_1(\mathbf{X},\mathbf{Y},\mathbf{Z}) = g(\mathbf{X}, \mathbf{Y})\ h(\mathbf{Y}, \mathbf{Z})$，其最大值点为

$$(\mathbf{x}_1^*, \mathbf{y}_1^*, \mathbf{z}_1^*) = \arg\max_{\mathbf{X},\mathbf{Y},\mathbf{Z}} F_1(\mathbf{X},\mathbf{Y},\mathbf{Z}).$$

再设 $F_2(\mathbf{X},\mathbf{Y}) = g(\mathbf{X}, \mathbf{Y}) \max_{\mathbf{Z}} h(\mathbf{Y}, \mathbf{Z})$，故其最大值点为

$$(\mathbf{x}^*, \mathbf{y}^*) = \arg\max_{\mathbf{X},\mathbf{Y}} F_2(\mathbf{X},\mathbf{Y}).$$

考虑式 (4.9) 等号的右边，有

$$\begin{aligned}\max_{\mathbf{X},\mathbf{Y}} F_2(\mathbf{X},\mathbf{Y}) &\geqslant F_2(\mathbf{x}_1^*, \mathbf{y}_1^*)\\ &= g(\mathbf{x}_1^*, \mathbf{y}_1^*) \max_{\mathbf{Z}} h(\mathbf{y}_1^*, \mathbf{Z})\\ &\geqslant g(\mathbf{x}_1^*, \mathbf{y}_1^*)\ h(\mathbf{y}_1^*, \mathbf{z}_1^*)\\ &= \max_{\mathbf{X},\mathbf{Y},\mathbf{Z}} F_1(\mathbf{X},\mathbf{Y},\mathbf{Z}).\end{aligned}$$

反过来，考虑式 (4.9) 等号的左边，有

$$\max_{\mathbf{X},\mathbf{Y},\mathbf{Z}} F_1(\mathbf{X},\mathbf{Y},\mathbf{Z}) \geqslant \max_{\mathbf{Z}} F_1(\mathbf{x}^*, \mathbf{y}^*, \mathbf{Z})$$

$$=\max_{\mathbf{Z}} g(\mathbf{x}^*,\mathbf{y}^*)\ h(\mathbf{y}^*,\mathbf{Z})$$

$$=\max_{\mathbf{X},\mathbf{Y}} F_2(\mathbf{X},\mathbf{Y}).$$

所以式（4.9）成立. 同时，$\mathbf{x}_1^* = \mathbf{x}^*$，$\mathbf{y}_1^* = \mathbf{y}^*$. 对于 $\mathbf{z}_1^*$，有

$$\begin{aligned}\mathbf{z}_1^* &= \arg\max_{\mathbf{Z}} g(\mathbf{x}^*,\mathbf{y}^*)\ h(\mathbf{y}^*,\mathbf{Z})\\ &= g(\mathbf{x}^*,\mathbf{y}^*)\arg\max_{\mathbf{Z}} h(\mathbf{y}^*,\mathbf{Z})\\ &= \arg\max_{\mathbf{Z}} h(\mathbf{y}^*,\mathbf{Z}) = \delta(\mathbf{y}^*).\end{aligned}$$

所以式（4.10）也成立. 定理得证. □

4.6.2 分解与计算复杂度

定理 4.5 提供了利用分解简化 NDP 问题的方法. 设 $f(\mathbf{H}) = g(\mathbf{X},\mathbf{Y})\ h(\mathbf{Y},\mathbf{Z})$. 根据定理 4.5，函数 $f(\mathbf{H})$ 的最大值可以通过如下两步来获得：

(1) $h'(\mathbf{Y}) = \max_{\mathbf{Z}} h(\mathbf{Y},\mathbf{Z})$；

(2) $\max_{\mathbf{X},\mathbf{Y}} g(\mathbf{X},\mathbf{Y})h'(\mathbf{Y})$.

这是否降低了计算复杂度呢？回答是肯定的. 设 $\mathbf{X},\mathbf{Y},\mathbf{Z}$ 各包含一个取 3 个值的变量：X,Y,Z. 那么，直接计算 $\max_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ 需要进行 $3^3=27$ 次数字乘法和 $3^3-1=26$ 次数字比较. 如果使用新方法，第一步对每个 Y 的取值 y 需要做 $3-1=2$ 次比较来计算函数 $h(y, Z)$ 的最大值，所以总共要做 $3\times 2=6$ 次比较；第二步需要 $3^2=9$ 次乘法和 $3^2-1=8$ 次比较. 比起直接计算，新方法节省了 18 次数字乘法和 12 次数字比较.

根据定理 4.5，函数 $f(\mathbf{H})$ 的最大值点可以通过如下步骤来获得：

(1) $\delta(\mathbf{Y}) = \arg\max_{\mathbf{Z}} h(\mathbf{Y},\mathbf{Z})$；

(2) $(\mathbf{x}^*, \mathbf{y}^*) = \arg\max_{\mathbf{X},\mathbf{Y}} g(\mathbf{X},\mathbf{Y})h'(\mathbf{Y})$；

(3) $\mathbf{z}^* = \delta(\mathbf{y}^*)$.

这个过程与计算最大值的过程做相同的数字运算（比较），不同的是要对 $\delta(\mathbf{Y})$ 进行记录. 因此，两个过程可以合二为一. 这意味着利用分解，我们可以同时降低最大值计算和最大值点计算的复杂度.

4.6.3 变量消元 MAP 算法

设 $\mathscr{F}=\{f_1, \cdots, f_m\}$ 是函数 $f(\mathbf{H})$ 的一个分解，ρ 是一个消元顺序. 图 4.15 给出了一个利用分解 $\mathscr{F}$ 计算 $(\max_{\mathbf{H}}, \arg\max_{\mathbf{H}} f(\mathbf{H}))$ 的算法，即变量消元 MAP 算法 VE-MAP. 设 Z 是 ρ 中第一个变量. VE-MAP 首先从 $\mathscr{F}$ 中找出所有涉及 Z 的因子 $f_1,\cdots,f_k$，把它们乘在一起，得到函数 $h(\mathbf{Y},Z) = \prod_{i=1}^{k} f_i$，这里 $\mathbf{Y}$ 是所有

$h(\mathbf{Y},Z)$ 所涉及的除 Z 之外的变量. 记 $\mathscr{F}$ 中所有其它因子的乘积为函数 $g(\mathbf{X},\mathbf{Y})$，则 $f(\mathbf{H})=g(\mathbf{X},\mathbf{Y})\,h(\mathbf{Y},Z)$，这里 $\mathbf{X}=\mathbf{H}\setminus(\mathbf{Y}\{Z\})$.

VE-MAP($\mathscr{F},\rho$)

输入：$\mathscr{F}$—— 函数 $f(\mathbf{H})$的一个分解；

　　ρ —— 包含 $\mathbf{H}$ 中所有变量的消元顺序.

输出：$(\max\limits_{\mathbf{H}} f(\mathbf{H}),\arg\max\limits_{\mathbf{H}} f(\mathbf{H}))$.

1：**if**(ρ 只含一个变量 Z)

2：　　将 $\mathscr{F}$ 中所有因子相乘得到 $f(Z)$；

3：　　**return**$(\max\limits_{Z} f(Z),\arg\max\limits_{Z} f(Z))$.

4：**end if**

5：从 ρ 中删去第一个变量 Z；

6：从 $\mathscr{F}$ 中删去所有涉及 Z 的函数 $f_1,\cdots,f_k$；

7：$h\leftarrow\prod\limits_{i=1}^{k} f_i$；

8：$\delta(\mathbf{Y})\leftarrow\arg\max\limits_{Z} h(\mathbf{Y},Z)$，其中 $\mathbf{Y}$ 为 h 中除 Z 以外的所有变量，$\mathbf{X}=\mathbf{H}\setminus(\mathbf{Y}\cup\{Z\})$；

9：$h(\mathbf{Y})\leftarrow\max\limits_{Z} h(\mathbf{Y},Z)$，$\mathscr{F}\leftarrow\mathscr{F}\cup(h)$；

10：$(v,\mathbf{x}^*,\mathbf{y}^*)\leftarrow$VE-MAP($\mathscr{F},\rho$)；　//递归调用

11：**return**$(v,(\mathbf{y}^*,\delta(\mathbf{y}^*)))$.

图 4.15　算法 VE-MAP：利用分解计算最大后验假设及其概率

根据式 (4.9) 和式 (4.10)，有

$$\max_{\mathbf{H}} f(\mathbf{H})=\max_{\mathbf{X},\mathbf{Y}}\left[g(\mathbf{X},\mathbf{Y})\max_{Z} h(\mathbf{Y},Z)\right] \tag{4.11}$$

$$\arg\max_{\mathbf{H}} f(\mathbf{H})=(\mathbf{x}^*,\mathbf{y}^*,\delta(\mathbf{y}^*)). \tag{4.12}$$

这里，

$$\delta(\mathbf{Y})=\arg\max_{Z} h(\mathbf{Y},Z), \tag{4.13}$$

$$(\mathbf{x}^*,\mathbf{y}^*)=\arg\max_{\mathbf{X},\mathbf{Y}}\left[g(\mathbf{X},\mathbf{Y})\max_{Z} h(\mathbf{Y},Z)\right]. \tag{4.14}$$

如图 4.15 所示，VE-MAP 的第 8、9 两行计算 $\arg\max\limits_{Z} h(\mathbf{Y},Z)$和$\max\limits_{Z} h(\mathbf{Y},Z)$第 10 行通过递归调用实现式 (4.11) 和式 (4.14)，而第 11 行则利用式 (4.12) 构造最后结果. 第 1～4 行处理只剩一个变量未消去的情况.

例 4.11　对如图 4.4 所示的贝叶斯网，求 $\max\limits_{A,B,C,D,E} P(A,B,C,D,E\mid F=0)$ 和 $\arg\max\limits_{A,B,C,D,E} P(A,B,C,D,E\mid F=0)$.

函数 $P(A,B,C,D,E,F=0)$ 的分解为

$$\mathscr{F}=\{P(A),\ P(B),\ P(C),\ P(D\mid A,B),\ P(E\mid B,C),\ P(F=0\mid D,E)\}.$$

设消元顺序为 $\rho = \langle C, E, B, D, A \rangle$，VE-MAP 的运算过程如下.

(1) 首先消去变量 C：

$$\delta_1(B, E) = \arg\max_C P(C)P(E \mid B, C),$$

$$\mathscr{F} = \{P(A), P(B), P(D \mid A, B), P(F = 0 \mid D, E), \psi_1(B, E)\},$$

其中 $\psi_1(B, E) = \max_C P(C)P(E \mid B, C)$.

(2) 递归调用 (1)，消去变量 E：

$$\delta_2(B, D) = \arg\max_E P(F = 0 \mid D, E)\psi_1(B, E),$$

$$\mathscr{F} = \{P(A), P(B), P(D \mid A, B), \psi_2(B, D)\},$$

其中 $\psi_2(B, D) = \max_E P(F = 0 \mid D, E)\psi_1(B, E)$.

(3) 递归调用 (2)，消去变量 B：

$$\delta_3(A, D) = \arg\max_B P(B)\ P(D \mid A, B)\psi_2(B, D),$$

$$\mathscr{F} = \{P(A), \psi_3(A, D)\},$$

其中 $\psi_3(A, D) = \max_B P(B)\ P(D \mid A, B)\psi_2(B, D)$.

(4) 递归调用 (3)，消去变量 D：

$$\delta_4(A) = \arg\max_D \psi_3(A, D),$$

$$\mathscr{F} = \{P(A), \psi_4(A)\},$$

其中 $\psi_4(A) = \max_D \psi_3(A, D)$.

(5) 递归调用 (4)，消去变量 A：因为 $\mathscr{F}$ 只剩一个变量，所以直接进入第 (2) 步，返回 (v, a^*)：

$$v = \max_A P(A)\psi_4(A),$$

$$a^* = \arg\max_A P(A)\psi_4(A).$$

(6) 接下来是一连串的递归调用返回：

递归调用 (3) 返回 $(v, (a^*, d^*))$，这里 $d^* = \delta_4(a^*)$；

递归调用 (2) 返回 $(v, (a^*, d^*, b^*))$，这里 $b^* = \delta_3(a^*, d^*)$；

递归调用 (1) 返回 $(v, (a^*, d^*, b^*, e^*))$，这里 $e^* = \delta_2(b^*, d^*)$.

(7) 最后主进程返回 $(v, (a^*, d^*, b^*, e^*, c^*))$，这里 $c^* = \delta_1(b^*, e^*)$.

最后算法返回 $(v, (a^*, d^*, b^*, e^*, c^*))$，其中，

$$(a^*, d^*, b^*, e^*, c^*) = \arg\max_{A,B,C,D,E} P(A,B,C,D,E,F = 0),$$

$$v = \max_{A, B, C, D, E} P(A, B, C, D, E, F = 0).$$

因此我们所求的值分别为

$$\arg\max_{A, B, C, D, E} P(A, B, C, D,E \mid F = 0) = (a^*, b^*, c^*, d^*, e^*),$$

$$\max_{A, B, C, D, E} P(A, B, C, D, E \mid F = 0) = \frac{v}{P(F = 0)}.$$

□

4.7 文献介绍

贝叶斯网推理的 3 个主要问题：后验概率问题、最大后验假设问题（MAP）和最大可能解释问题（MPE），都是 NP-难解的（Cooper，1990；Shimony，1994；Park，2002）. MPE 是典型的组合优化问题，它的判定问题是 NP-完全的（Shimony，1994；Park，2002）；后验概率问题是一个计数问题（counting problem），它是 #P-完全的（Littman et al.，2001；Roth，1996）；MAP 是 NP^{PP}-完全的（Park，2002）.

变量消元法是由 Zhang 和 Poole（1994，1996）首先提出的，其原理源自关于不定序动态规划的研究（Bertele and Brioschi，1972）. 相近的工作包括 D'Ambrosio（1991）、Shachter 等（1994）、Shenoy（1992）和 Dechter（1999）等人的研究. 变量消元法的变种也曾独立地出现于其它领域，如计算生物学中进行家系树分析的“pruning-and-peeling”算法（Cannings，1978）、隐马尔可夫模型中的“forward-backward”算法（Rabiner and Juang，1993），以及状态空间模型中的“Kalman filtering-smoothing”算法（Shalom and Li，1993）等.

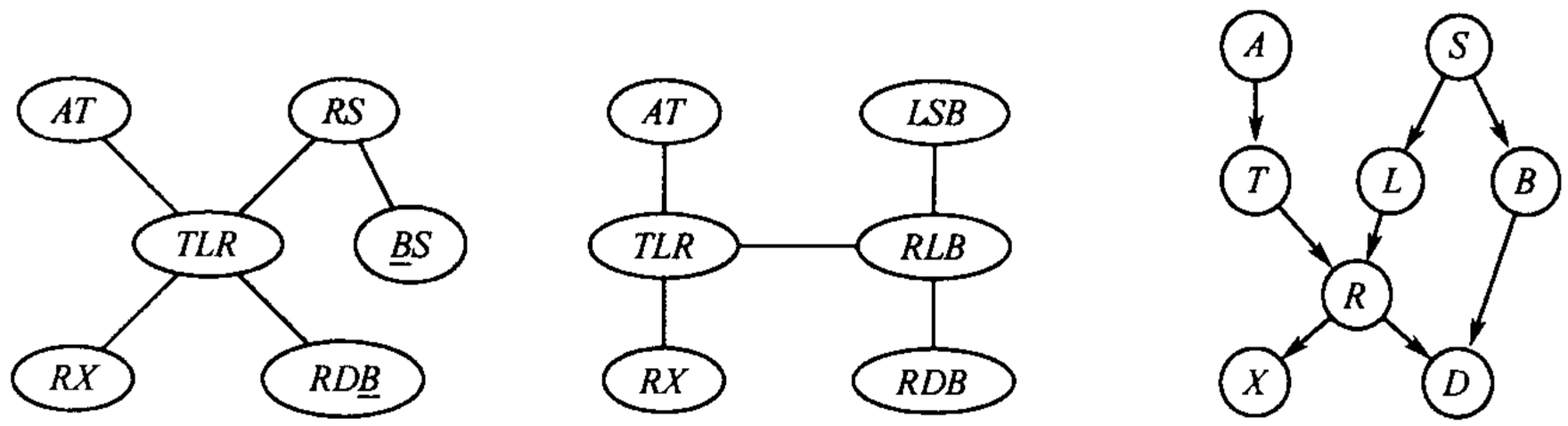
最优消元顺序问题是 NP-难解的（Arnborg et al.，1987）. 最大势搜索和最小缺边搜索算法分别由 Tarjan 和 Yannakakis（1984）以及 Bertele 和 Brioschi（1972）给出. 近期关于最优消元顺序的研究可参见有关文献（Becker and Geiger，2001；Lauritzen and Jensen，1997；Meila and Jordan，1997a，b）.

第 5 章　团树传播算法

在实际应用中，往往需要在同一贝叶斯网中进行多次不同的推理，而两次不同的推理之间往往存在一些相同的步骤. 变量消元法逐一处理推理问题，不考虑步骤共享. 本章介绍一个利用步骤共享来加快推理的算法，即团树传播法. 它能用大约两倍于变量消元法计算一个变量的后验概率的时间，计算出网络中每个变量的后验概率. 下面首先引入团树的概念，然后介绍如何在团树中进行信息传播，实现计算共享，最后讨论团树的构造等问题.

5.1　团　　树

团树（clique tree）是一种无向树，其中每一个节点代表一个变量集合，称为**团**（clique）. 团树必须满足**变量连通性**（variable-connectedness），即包含同一变量的所有团所导出的子图必须是连通的. 换句话说，如果团树中的两个团 $\mathbf{C}_1$ 和 $\mathbf{C}_2$ 同时包含某变量 X，那么在连接 $\mathbf{C}_1$ 和 $\mathbf{C}_2$ 的通路上的所有团都必须包含 X. 图 5.1(a)所示的树不是团树，因为变量 B 既在团 $[BS]$ ①中，又在团 $[RDB]$ 中，但却不在两者之间的通路上的团 $[TLR]$ 和 $[RS]$ 中. 图 5.1(b)所示的树是变量连通的，因而是团树.



(a) 非团树　　(b) 团树 $\mathcal{T}$　　(c) 贝叶斯网 $\mathcal{N}$

图 5.1　团树与贝叶斯网

我们说团树 $\mathcal{T}$ **覆盖**（cover）贝叶斯网 $\mathcal{N}$，如果它满足以下两个条件：

① 为简便起见，我们有时把节点集合 $\{B, S\}$ 记为 $[BS]$ 的形式.

(1) 对 $\mathcal{N}$ 中任一变量 X，在 $\mathcal{T}$ 中有一个团 $\mathbf{C}$，使得 $X \in \mathbf{C}$ 且 $\pi(X) \subseteq \mathbf{C}$；

(2) $\mathcal{T}$ 中所有团的并集刚好是 $\mathcal{N}$ 中所有变量的集合.

图 5.1(b)中的团树覆盖图 5.1(c)中的贝叶斯网. 比如，变量 R 和它的父节点 T，L 都在团 $[TLR]$ 中，变量 D 和它的父节点 R，B 都在团 $[RDB]$ 中，等等. 条件 (1) 中的团 $\mathbf{C}$ 称为变量 X 的**家族覆盖团** (cover clique). 在图 5.1(b)中，$[LSB]$ 是变量 L 和 B 的家族覆盖团，$[RDB]$ 是变量 D 的家族覆盖团，但 $[RLB]$ 不是 $\mathcal{N}$ 中任何变量的家族覆盖团.

第 5.6 节将会介绍怎样构造覆盖一个贝叶斯网的团树，下面先讨论怎样利用这样的团树来组织贝叶斯网推理.

5.2　一个变量后验概率的计算

设 $\mathcal{N}$ 为一个贝叶斯网，$\mathcal{T}$ 是覆盖 $\mathcal{N}$ 的一棵团树，$\mathbf{E} = \mathbf{e}$ 是观测到的证据，Q 是一个查询变量. 图 5.2 给出了一个利用 $\mathcal{T}$ 来计算后验概率 $P(Q \mid \mathbf{E} = \mathbf{e})$ 的算法 CTP_0. 本节逐步解释这个算法，下一节证明它的正确性.

图 5.2 为 CTP_0 算法示意图. CTP_0 的第 1 步是利用 $\mathcal{N}$ 中的概率函数将团树 $\mathcal{T}$ 初始化，即将 $\mathcal{N}$ 中的概率函数分配到 $\mathcal{T}$ 中各节点加以储存. 对 $\mathcal{N}$ 中的每一变量 X，在 $\mathcal{T}$ 中找到它的一个家族覆盖团，然后把 $P(X \mid \pi(X))$ 储存于该团处. 在分配过程结束后，如果有团没有储存任何概率分布，则构造单位函数 $= \mathbf{1}$，存于该团处. 作为一个贯穿本节的例子，考虑如图 5.1(b)所示的团树 $\mathcal{T}$ 和图 5.1(c)所示的贝叶斯网 $\mathcal{N}$. 这里对 $\mathcal{T}$ 初始化 的结果如图 5.3(a)所示.

CTP_0 的第 2 步是设置证据. 在我们的例子中，假设已知证据 $\{D = y\}$，设置证据后的团树 $\mathcal{T}$ 如图 5.3(b)所示. 团 $[RDB]$ 所储存的函数由 $P(D \mid R,B)$ 变成了 $P(D = y \mid R,B)$.

CTP_0 的第 3 步是要选择一个包含查询变量 Q 的团 $\mathbf{C}_Q$ 作为推理的枢纽节点，注意 $\mathbf{C}_Q$ 不一定是 Q 的家族覆盖团. 在我们的例子中，假设要计算后验概率 $P(T \mid D = y)$，那么既可以用团 $[AT]$，也可以用团 $[TLR]$ 作为推理枢纽. 下面假设以 $[TLR]$ 为枢纽.

在第 4 步，CTP_0 对 $\mathbf{C}_Q$ 的相邻节点逐一调用子程序 $\mathrm{CollectMessage}_0(\mathcal{T}, \mathbf{C}_Q, \mathbf{C})$，并获得一个函数. 为了便于直观理解，从相邻节点 $\mathbf{C}$ 那里获得的函数称为是从 $\mathbf{C}$ 到 $\mathbf{C}_Q$ 的**信息**. 一般地，称对 $\mathrm{CollectMessage}_0(\mathcal{T}, \mathbf{C}', \mathbf{C})$ 的调用是 $\mathbf{C}'$ 从 $\mathbf{C}$ **获得信息**，或 $\mathbf{C}$ 向 $\mathbf{C}'$ **传递信息**. 在我们的例子中，枢纽节点有 3 个邻居：$[AT]$、$[RX]$ 和 $[RLB]$. 下面是计算从它们到 $[TLR]$ 的信息的过程：

(1) 从 $[AT]$ 到 $[TLR]$ 的信息：储存在 $[AT]$ 处的函数为 $P(A)$ 和 $P(T \mid A)$，集合 $\mathbf{Z} = \{A\}$. 因为 $[AT]$ 在 $\mathcal{T}$ 中是叶节点，所以从 $[AT]$ 到 $[TLR]$

的信息为如下函数：

$$\psi_1(T) = \sum_{A} P(A)P(T \mid A),$$

CTP$_0$ ($\mathcal{N}$, $\mathcal{T}$, **E**, **e**, Q)

输入： $\mathcal{N}$——一个贝叶斯网； $\mathcal{T}$——一棵覆盖 $\mathcal{N}$ 的团树；
E——证据变量； **e**——证据变量的取值；
Q——一个查询变量.

输出： $P(Q \mid \mathbf{E} = \mathbf{e})$；

1：利用 $\mathcal{N}$ 将 $\mathcal{T}$ 初始化，即将 $\mathcal{N}$ 中的概率分布函数存储于 $\mathcal{T}$ 中的各个节点处；
2：在 $\mathcal{T}$ 的函数中，将证据变量 **E** 设置为其取值 **e**；
3：在 $\mathcal{T}$ 中找出一个包含 Q 的团 $\mathbf{C}_Q$ 作为枢纽；
4：**for**（每个与 $\mathbf{C}_Q$ 相邻的节点 **C**）
5：　　调用 CollectMessage$_0$ ($\mathcal{T}$, $\mathbf{C}_Q$, **C**) 获得一个函数；
6：**end for**
7：把上一步所获得函数以及储存在 $\mathbf{C}_Q$ 处的函数相乘，得到 $\mathbf{C}'_Q$ 的一个函数 $h(\mathbf{C}'_Q)$，这里 $\mathbf{C}'_Q = \mathbf{C}'_Q \setminus \mathbf{E}$；
8：**return** $\sum_{\mathbf{C}'_Q \setminus \{Q\}} h(\mathbf{C}'_Q) \Big/ \sum_{\mathbf{C}'_Q} h(\mathbf{C}'_Q)$.

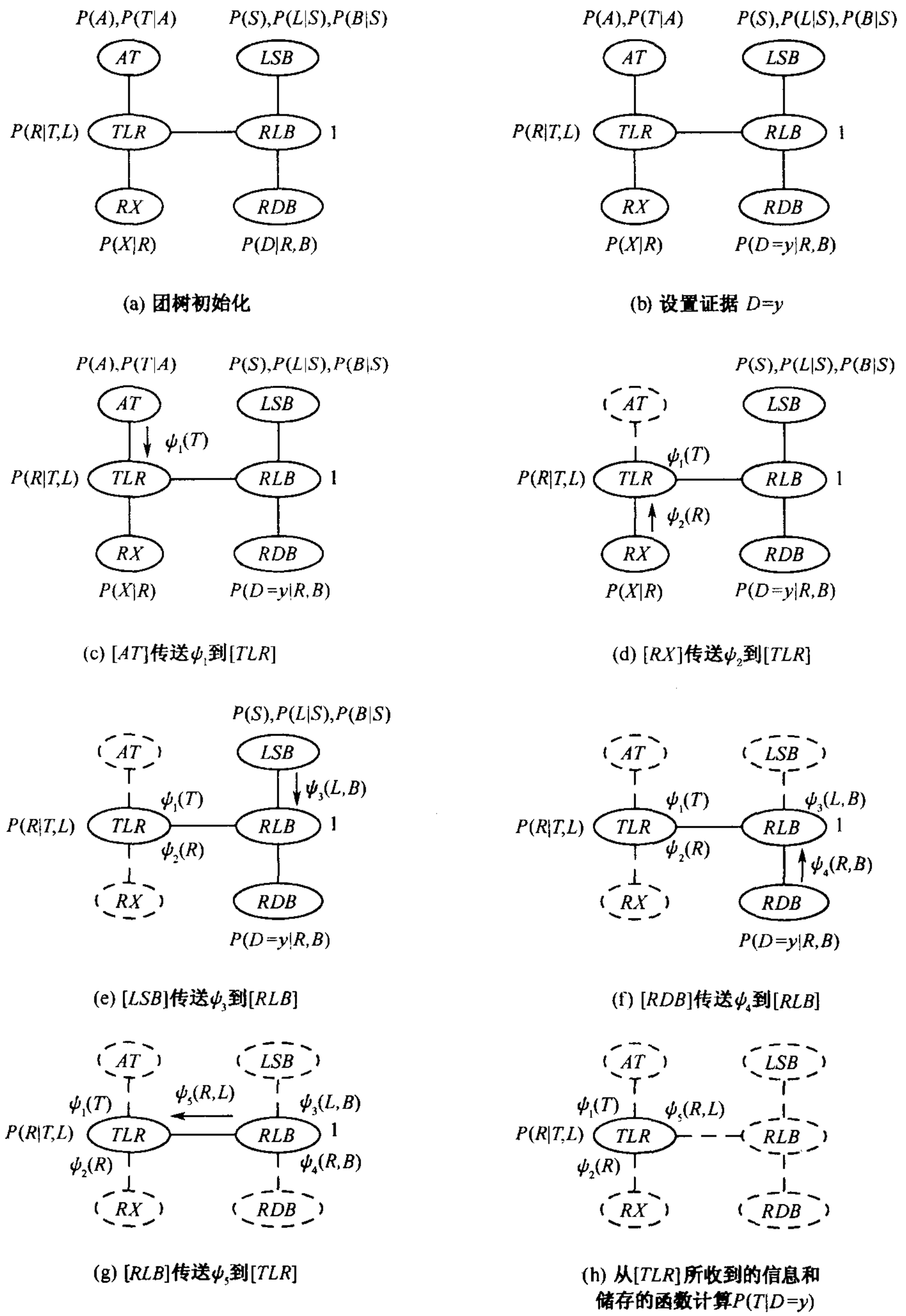
CollectMessage$_0$ ($\mathcal{T}$, $\mathbf{C}'$, **C**)

输入： $\mathcal{T}$——一棵初始化了的覆盖 $\mathcal{N}$ 的团树；
$\mathbf{C}'$, **C**——团树中两个相邻的节点.

输出： 一个函数（从 **C** 到 $\mathbf{C}'$ 的信息）；

1：设 f_1, f_2, …, f_l 为储存在 **C** 处的函数，且设 $\mathbf{Z} = \mathbf{C} \setminus (\mathbf{C}' \cup \mathbf{E})$；
2：**if**（**C** 为 $\mathcal{T}$ 中的叶节点）
3：　　**return** $\sum_{\mathbf{Z}} \prod_{i=1}^{l} f_i$.
4：**else**
5：　　设 $\mathbf{C}'_1$, $\mathbf{C}'_2$, …, $\mathbf{C}'_k$ 为除 $\mathbf{C}'$ 以外的其它与 **C** 相邻的节点；
6：　　**for** $j = 1$ **to** k
7：　　$g_j \leftarrow$ CollectMessage$_0$ ($\mathcal{T}$, **C**, $\mathbf{C}'_i$)；
8：　　**end for**
9：　　**return** $\sum_{\mathbf{Z}} \prod_{i=1}^{l} f_i \prod_{j=1}^{k} g_j$.
10：**end if**

图 5.2　CTP$_0$：单查询变量团树传播算法



(a) 团树初始化

(b) 设置证据 $D=y$

(c) $[AT]$传送ψ_1到$[TLR]$

(d) $[RX]$传送ψ_2到$[TLR]$

(e) $[LSB]$传送ψ_3到$[RLB]$

(f) $[RDB]$传送ψ_4到$[RLB]$

(g) $[RLB]$传送ψ_5到$[TLR]$

(h) 从$[TLR]$所收到的信息和储存的函数计算$P(T|D=y)$

图 5.3　单查询变量团树传播示例

这一步如图 5.3(c)所示.

(2) 从 [RX] 到 [TLR] 的信息：与上一步相似，[TLR] 从 [RX] 处获得的信息为如下函数：

$$\psi_2(R) = \sum_X P(X \mid R),$$

这一步如图 5.3(d)所示.

(3) 从 [RLB] 到 [TLR] 的信息：由于除 [TLR] 以外，[RLB] 还有其它两个邻居 [LSB] 和 [RDB]，所以在计算从 [RLB] 到 [TLR] 的信息之前，需要先计算从 [LSB] 和 [RDB] 到 [RLB] 的信息，如图 5.3(e) 和 (f) 所示，它们分别为

$$\psi_3(L,B) = \sum_S P(S)P(L \mid S)P(B \mid S),$$

$$\psi_4(R,B) = P(D = y \mid R,B),$$

注意在计算 ψ_4 时，因为 D 是证据变量，所以 $\mathbf{Z}$ 是空集. 最后，从 [RLB] 到 [TLR] 的信息为如下函数：

$$\psi_5(R, L) = \sum_B \psi_3(L,B) \cdot \psi_4(R,B),$$

这一步如图 5.3(g)所示.

在第 5 步，CTP_0 把枢纽节点 $\mathbf{C}_Q$ 所收到的信息及储存在 $\mathbf{C}_Q$ 处的函数相乘，得到一个 $\mathbf{C}'_Q$ 的函数 $h(\mathbf{C}'_Q)$. 在我们的例子中，这个函数为

$$h(T,L,R) = \psi_1(T)\psi_2(R)\psi_5(R,L)P(R \mid T,L),$$

如图 5.3(h)所示.

在第 6 步，CTP_0 从 $h(\mathbf{C}'_Q)$ 中消去除 Q 以外的变量，并将结果归一化，从而得到 $P(Q \mid \mathbf{E} = \mathbf{e})$. 在我们的例子中，即

$$P(T \mid D = y) = \frac{\sum_{L,R} h(T,L,R)}{\sum_{T,L,R} h(T,L,R)}.$$

枢纽团 [TLR] 中的另外两个变量 L 和 R 的后验分布 $P(L \mid D = y)$ 和 $P(R \mid D = y)$ 也可以从 $h(T,L,R)$ 中用同样方法得到.

CTP_0 可以视为是用团树来组织变量消元的算法. 在上面的例子中，团 [AT] 向团 [TLR] 传递消息 $\psi_1(T)$ 实际上就是消去变量 A；[RX] 向 [TLR] 传递消息 $\psi_2(R)$ 是消去变量 X；[LSB] 向 [RLB] 传递 $\psi_3(L,B)$ 是消去变量 S；[RLB] 向 [TLR] 传递 $\psi_5(R,L)$ 是消去变量 B. 不过，团树中的信息传递并不总代表消去一个变量. 在上面的例子中，[RDB] 向 [RLB] 传递信息 $\psi_4(R, B)$ 就不涉及变量消元. 另一方面，有时一次信息传递也可能对应于同时消去多个变量.

5.3　团树传播的正确性

考虑团树 $\mathcal{T}$ 中的一个团 **C**. 设 $\mathbf{C}'$ 是从 **C** 到 $\mathbf{C}_Q$ 的通路上与 **C** 相邻的节点，$\mathbf{C}'_1,\cdots,\mathbf{C}'_k$ 为其它与 **C** 相邻的节点，如图 5.4 所示. 在算法 CTP_0 的运行过程中，$\mathrm{CollectMessage}_0(\mathcal{T},\mathbf{C}',\mathbf{C})$ 会被调用一次，而且仅仅被调用一次. 换句话说，**C** 会向 $\mathbf{C}'$ 传递一次信息，而且仅仅传递一次. 在 **C** 向 $\mathbf{C}'$ 传递信息之前，**C** 称为是**未被激活的**（inactivated），而在 **C** 向 $\mathbf{C}'$ 传递信息之后，**C** 称为是**已被激活的**（activated）. 因此，算法 CTP_0 的运行过程可以看做是一个 $\mathcal{T}$ 中的节点逐个被激活的过程. 在激活 $\mathbf{C}'_1,\cdots,\mathbf{C}'_k$ 之后，节点 **C** 收到信息 $g_1,\cdots,g_k$，基于这些信息以及自身存储的函数 $f_1,\cdots,f_l$，节点 **C** 计算新的信息 ψ 并将其传递给 $\mathbf{C}'$，之后 **C** 变成是已被激活的.

命题 5.1　在算法 CTP_0 第 4 步的运行过程中，下面两个性质始终保持不变：

（1）对任何一个团 **C**，储存在、以及传递到 **C** 处的所有函数中所涉及的变量全部都在 **C** 中；

（2）储存在、以及传递到所有未被激活的团处的函数的乘积是 $P(\mathbf{X},\mathbf{E}=\mathbf{e})$，其中 **X** 是未被激活的团中所有非证据变量的集合.

证明　采用归纳法. 考虑某个变量 X 的概率分布函数 $P(X\mid\pi(X))$，设它储存在团 **C** 处. 在第 1 步初始化时，$P(X\mid\pi(X))$ 之所以被分配到团 **C** 处是因为 **C** 是 X 的一个家族覆盖团，即有 $X\in\mathbf{C}$ 和 $\pi(X)\subseteq\mathbf{C}$. 所以，$P(X\mid\pi(X))$ 中的所有变量都在 **C** 中，这就证明了在第 4 步信息传递开始之前，性质（1）成立. 另一方面，在第 1 步初始化之后，$\mathcal{T}$ 中所有函数的乘积是贝叶斯网 $\mathcal{N}$ 中 所有变量的联合分布 $P(\mathbf{X},\mathbf{E})$. 在第 2 步设置证据之后，这个乘积变成 $P(\mathbf{X},\mathbf{E}=\mathbf{e})$. 所以，在信息传递开始之前，性质（2）也成立.

考虑图 5.4 中团 **C** 被激活的过程. 假设 $\mathbf{C}'_1,\cdots,\mathbf{C}'_k$ 已被激活、而团 **C** 未被激活时，性质（1）和（2）都成立. 下面证明在 **C** 被激活后它们仍然成立.

从 **C** 到 $\mathbf{C}'$ 的信息由 $\mathrm{CollectMessage}_0$ 第 9 行给出，其中 f_i 是储存在 **C** 处的本地函数，g_j 从其它团传递到 **C** 处的函数. 根据归纳假设，f_i 和 g_j 中的所有变量都在 **C** 中，因而在 $\mathbf{C}\setminus\mathbf{E}$ 中. 因为 **Z** 是所有在 **C** 中而不在 $\mathbf{C}'\cup\mathbf{E}$ 中的变量，所以 ψ 中的所有变量都在 $\mathbf{C}'$ 中. 因此，性质（1）在信息传递过程中保持不变.

设 **X** 是在 $\mathbf{C}'_1,\cdots,\mathbf{C}'_k$ 已被激活，而团 **C** 未被激活时所有未被激活的团中的非证据变量的集合，$\mathbf{X}'$ 是 **C** 被激活后所有未被激活的团中的非证据变量的集合. 进一步设 r 是 **C** 被激活后，除 ψ 之外的、存储在或传递到所有未被激活的团处的函数的乘积. 根据已证明的性质（1），r 所涉及的变量全部在 $\mathbf{X}'$ 中，所以可以将 r 写成 $r(\mathbf{X}')$.

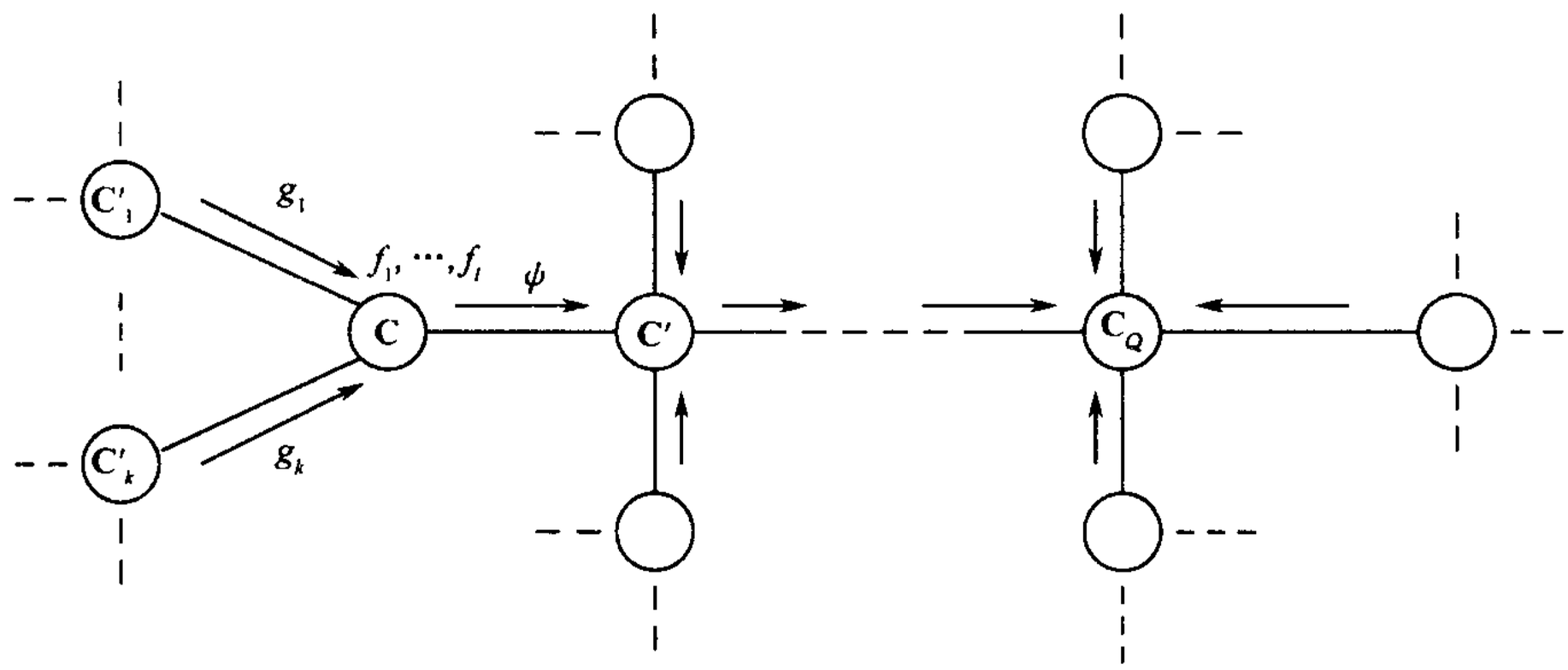


图 5.4　信息从叶节点传向枢纽节点 $\mathbf{C}_Q$

注意，r 同样也是在 $\mathbf{C}$ 向 $\mathbf{C}'$ 传递信息之前，所有除了 f_i 及 g_j 之外的储存在或传递到未被激活的团处的函数的乘积. 按照归纳假设，有

$$r(\mathbf{X}')\prod_{i=1}^{l} f_i \prod_{j=1}^{k} g_j = P(\mathbf{X}, \mathbf{E}=\mathbf{e}) . \tag{5.1}$$

接下来考虑集合 $\mathbf{Z}=\mathbf{C}\setminus(\mathbf{C}'\cup\mathbf{E})$. 显然，$\mathbf{Z}\cup\mathbf{X}'=\mathbf{X}$. 对任一 $Z\in\mathbf{Z}$，有 $Z\in\mathbf{C}$，但 $Z\notin\mathbf{C}'$. 由于 $\mathcal{T}$ 的变量连通性，必有 $Z\notin\mathbf{X}'$，从而 $Z\cap\mathbf{X}'=\varnothing$. 于是，

$$\begin{aligned}
r(\mathbf{X}')\cdot\psi &= r(\mathbf{X}')\sum_{\mathbf{Z}}\prod_{i=1}^{l} f_i \prod_{j=1}^{k} g_j \\
&= \sum_{\mathbf{Z}} r(\mathbf{X}')\prod_{i=1}^{l} f_i \prod_{j=1}^{k} g_j \quad (\text{因为 } \mathbf{Z}\cap\mathbf{X}'=\varnothing) \\
&= \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{E}=\mathbf{e}) \quad (\text{因为式}(5.1)) \\
&= \sum_{\mathbf{Z}} P(\mathbf{Z}, \mathbf{X}', \mathbf{E}=\mathbf{e}) \quad (\text{因为 } \mathbf{X}=\mathbf{X}'\cup\mathbf{Z},\ \mathbf{X}'\cap\mathbf{Z}=\varnothing) \\
&= P(\mathbf{X}', \mathbf{E}=\mathbf{e}) .
\end{aligned}$$

这就证明了在 $\mathbf{C}$ 被激活之后，性质（2）仍然成立. 命题得证. □

定理 5.1　算法 $\mathrm{CTP}_0(\mathcal{N}, \mathcal{T}, \mathbf{E}, \mathbf{e}, Q)$ 返回的是变量 Q 的后验分布 $P(Q\mid\mathbf{E}=\mathbf{e})$.

证明　根据命题 5.1，CTP_0 在第 5 步所获得的函数 h 其实是 $P(\mathbf{C}'_Q, \mathbf{E}=\mathbf{e})$. 所以第 6 步所返回的是

$$\frac{\sum_{\mathbf{C}'_Q\setminus\{Q\}} h(\mathbf{C}'_Q)}{\sum_{\mathbf{C}'_Q} h(\mathbf{C}'_Q)} = \frac{\sum_{\mathbf{C}'_Q\setminus\{Q\}} P(\mathbf{C}'_Q, \mathbf{E}=\mathbf{e})}{\sum_{\mathbf{C}'_Q} P(\mathbf{C}'_Q, \mathbf{E}=\mathbf{e})}$$

$$=\frac{P(Q,\mathbf{E}=\mathbf{e})}{P(\mathbf{E}=\mathbf{e})}$$
$$=P(Q\mid\mathbf{E}=\mathbf{e})\ .$$

定理得证. □

激活一个团的计算复杂度可以用 CollectMessage_0 的第 9 行所涉及的函数 $\prod_{i=1}^{l} f_i \prod_{j=1}^{k} g_j$ 的复杂度来度量. 根据命题 5.1，这个函数的所有变量都在团 **C** 中. 另一方面，如果团树是按照后面第 5.6 节中所介绍的方法构造的，而且 **C** 中无证据变量，那么 **C** 中所有变量都会出现在该函数中. 因此，**C** 向 **C**′ 传递信息的复杂度可以用下式来度量：

$$\prod_{X\in\mathbf{C}} |X|,$$

它是团 **C** 的大小 $|\mathbf{C}|$ 的一个指数函数. 算法 CTP_0 的总复杂度可以用下式来度量：

$$\sum_{\mathbf{C}\in\mathcal{T}}\prod_{X\in\mathbf{C}} |X|,$$

其中对应于最大团的那一项往往起主导作用. 所以，人们通常说概率推理的复杂度是最大团大小的指数函数.

5.4　团树传播与计算共享

CTP_0 算法与第 4 章介绍的变量消元算法 VE 并无本质区别，它是变量消元的另一种组织形式. 用团树组织变量消元的优点在于，它能使我们清楚地看到两次不同的推理计算之间哪些步骤是相同的，从而可以进行步骤共享，提高推理效率.

在图 5.3 中，为了计算变量 T 的后验分布 $P(T\mid D=y)$，进行了如下信息传递：$[AT]\to[TLR]$、$[RX]\to[TLR]$、$[LSB]\to[RLB]$、$[RDB]\to[RLB]$ 和 $[RLB]\to[TLR]$. 现假设需要进一步计算变量 B 的后验分布 $P(B\mid D=y)$. 如果选择团 $[RLB]$ 为枢纽节点，则需进行如下的消息传递：$[AT]\to[TLR]$、$[RX]\to[TLR]$、$[TLR]\to[RLB]$、$[LSB]\to[RLB]$ 和 $[RDB]\to[RLB]$. 这里除了信息 $[TLR]\to[RLB]$ 以外，所有其它信息传递都在计算 $P(T\mid D=y)$ 时进行过了. 所以，如果能在两次推理之间实现计算共享，原则上只需要再进行一次信息传递 $[TLR]\to[RLB]$，就可以完成计算 $P(B\mid D=y)$ 所需的信息传递.

要想在两次不同推理之间实现计算共享，就需要储存和利用中间结果. 为此，可以引入两个子程序 SaveMessage(·)和 RetrieveMessage(·). 设已计算出从团 **C** 到其相邻节点 **C**′ 的信息是函数 ψ，SaveMessage(**C**, **C**′, ψ) 把这个结果

CTP$_1$ ($\mathcal{N}$, $\mathcal{T}$, **E**, **e**, Q)

输入：$\mathcal{N}$——一个贝叶斯网；$\mathcal{T}$——一棵覆盖 $\mathcal{N}$ 的团树.

E——证据变量；**e**——证据变量的取值；

Q——一个查询变量.

输出：$P(Q \mid \mathbf{E}=\mathbf{e})$；

1：利用 $\mathcal{N}$ 将 $\mathcal{T}$ 初始化，即将 $\mathcal{N}$ 中的概率分布函数存储于 $\mathcal{T}$ 中的各节点处；

2：在 $\mathcal{T}$ 中的函数中，将证据变量 **E** 设置为其取值 **e**；

3：在 $\mathcal{T}$ 中找出一个包含 Q 的团 $\mathbf{C}_Q$ 作为枢纽；

4：**for**（每个与 $\mathbf{C}_Q$ 相邻的节点 **C**）

5：　调用 CollectMessage$_1$ ($\mathcal{T}$, $\mathbf{C}_Q$,**C**) 获得一个函数；

6：**end for**

7：把上一步所获得的函数以及储存在 $\mathbf{C}_Q$ 处的函数相乘，得到 $\mathbf{C}'_Q$ 的一个函数 $h(\mathbf{C}'_Q)$，这里 $\mathbf{C}'_Q = \mathbf{C}_Q \setminus \mathbf{E}$；

8：**return** $\sum_{\mathbf{C}'_Q \setminus \{Q\}} h(\mathbf{C}'_Q) \Big/ \sum_{\mathbf{C}'_Q} h(\mathbf{C}'_Q)$.

CollectMessage$_1$ ($\mathcal{T}$, $\mathbf{C}'$, **C**)

输入：$\mathcal{T}$——一棵初始化了的覆盖 $\mathcal{N}$ 的团树；

$\mathbf{C}'$, **C**——团树中的两个相邻的节点.

输出：一个函数（从 **C** 到 $\mathbf{C}'$ 的信息）；

1：$\psi \leftarrow$ RetrieveMessage($\mathbf{C}$,$\mathbf{C}'$)；

2：**if**($\psi \neq$ **null**)

3：　**return** ψ;

4：设 f_1, f_2, …, f_l 为储存在 **C** 处的函数，且设 $\mathbf{Z} = \mathbf{C} \setminus (\mathbf{C}' \cup \mathbf{E})$；

5：**if**（**C** 在 $\mathcal{T}$ 中为叶节点）

6：　$\psi \leftarrow \sum_{Z} \prod_{i=1}^{l} f_i$;

7：**else**

8：　设 $\mathbf{C}'_1$, $\mathbf{C}'_2$, …,$\mathbf{C}'_k$ 为除 $\mathbf{C}'$ 以外的其它与 **C** 相邻的节点；

9：　**for** i=1 **to** k

10：　　$g_i \leftarrow$ CollectMessage$_1$ (**C**, $\mathbf{C}'_i$)；

11：　**end for**

12：　$\psi \leftarrow \sum_{\mathbf{Z}} \prod_{i=1}^{l} f_i \prod_{j=1}^{k} g_j$;

13：　SaveMessage (**C**, $\mathbf{C}'$, ψ)；

14：**end if**

15：**return** ψ.

图 5.5　算法 CTP$_1$：CTP$_0$ 的改进，在不同推理之间实现中间结果共享

储存起来，以备以后用 RetrieveMessage(**C**, **C**′) 调用. 如果在信息尚未储存之前调用 RetrieveMessage(**C**, **C**′)，所得到的将是一空函数. SaveMessage 和 RetrieveMessage可以有多种不同的实现方法. 方法之一是给节点 **C**′ 分配空间，专门用来储存来自节点 **C** 的信息，SaveMessage(·)和 RetrieveMessage(·)分别执行该处信息的存储和查询功能. 另一种方法是在给边“**C**-**C**′”分配空间，来存储相应的信息.

基于以上思想，利用 SaveMessage 和 RetrieveMessage 把 CollectMessage_0 修改成如图 5.5 中所示的 CollectMessage_1. 主要改动有两点：第一，在计算从 **C** 到 **C**′ 的信息之前，CollectMessage_1 首先调用 RetrieveMessage(**C**, **C**′)，查询要计算的信息是否已经存在. 如果该信息之前曾被计算过并储存起来，那么 RetrieveMessage(**C**, **C**′) 就能将之找回，从而节省一次计算. 第二，在计算从 **C** 到 **C**′ 的信息 ψ 之后，调用 SaveMessage(**C**, **C**′,ψ) 将其储存起来，以备后用.

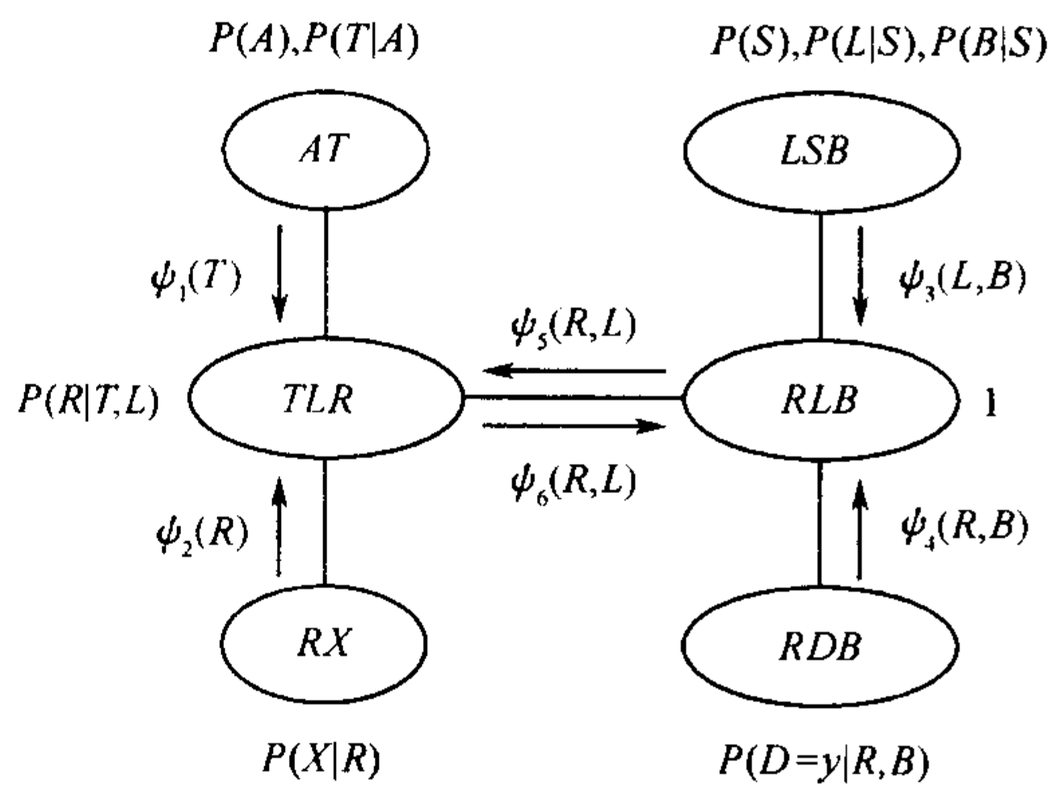


图 5.6　图树传播与计算共享

第二次推理利用了第一次推理的如下中间结果：

$\psi_1(T)$、$\psi_2(R)$、$\psi_3(L,B)$、$\psi_4(R,B)$

设 CTP_1 为将 CTP_0 中的 CollectMessage_0 换成 CollectMessage_1 所得到的算法，如图 5.5 所示. 不难看出，如果用 CTP_1 进行多次推理，那么后面的推理计算就可以利用前面所获得的中间结果. 在本节开始时所描述的例子中，CTP_1 在计算 $P(T \mid D=y)$ 时会储存所计算的 5 个信息 ψ_1，ψ_2，ψ_3，ψ_4 和 ψ_5，如图 5.6 所示. 在接下来计算 $P(B \mid D=y)$ 时，它会从储存空间中找出已经计算过的信息 ψ_1，ψ_2，ψ_3 和 ψ_4. 唯一需要计算的 是从 $[TLB] \rightarrow [RLB]$ 的信息，即

$$\psi_6(R,L) = \sum_T \psi_1(T)\psi_2(R)P(R \mid T,L).$$

这样就节省了大部分的运算.

5.5　每个变量的后验概率的计算

设 $\mathcal{N}$ 是一个贝叶斯网，$\mathbf{E}=\mathbf{e}$ 是观测到的证据. **团树传播算法**（clique tree propagation）的目的是要计算每一个非证据变量的后验概率分布. 它所做的运算与对每一个非证据变量分别调用 CTP_1 所需做的运算完全一样. 不同的是，它

重新组织了运算，先做共享运算，再做非共享运算.

团树传播算法，简记为 CTP，由图 5.7 给出. 它首先用贝叶斯网中的概率函数将团树 $\mathcal{T}$ 初始化，并且设置证据. 然后，任选一个团 $\mathbf{C}_P$ 作为枢纽节点，进行信息传递. 信息传递分为收集和分发两个阶段. 在信息收集阶段，即第 4 步，

CTP ($\mathcal{N}$, $\mathcal{T}$, **E**, **e**)

输入：$\mathcal{N}$——一个贝叶斯网；$\mathcal{T}$——一棵覆盖 $\mathcal{N}$ 的团树；
E——证据变量；**e**——证据变量的取值.
输出：每一个非证据变量 X 的后验概率分布 $P(X \mid \mathbf{E}=\mathbf{e})$；

1：利用 $\mathcal{N}$ 将 $\mathcal{T}$ 初始化；
2：在 $\mathcal{T}$ 中的函数中，将证据变量 **E** 设置为其取值 **e**；
3：在 $\mathcal{T}$ 中任选一个团节点 $\mathbf{C}_P$ 作为枢纽；
4：**for**（每一个与 $\mathbf{C}_P$ 相邻的节点 **C**）
5：　　调用 CollectMessage($\mathbf{C}_P$,**C**)；// $\mathbf{C}_P \leftarrow \mathbf{C}$
6：**end for**
7：**for**（每一个 $\mathbf{C}_P$ 相邻的节点 **C**）
8：　　调用 DistributeMessage($\mathbf{C}_P$, **C**)；//$\mathbf{C}_P \rightarrow \mathbf{C}$
9：**end for**
10：**for**（$\mathcal{N}$ 中每一个非证据变量 X）
　10.1：选择一个包含 X 的团 $\mathbf{C}_X$；
　10.2：把初始化时和信息传递过程中储存在 $\mathbf{C}_X$ 处的函数相乘，得到一个 $\mathbf{C}'_X$ 的函数 $h(\mathbf{C}'_X)$，这里 $\mathbf{C}'_X = \mathbf{C}_X \setminus \mathbf{E}$；
　10.3：**return** $\sum_{\mathbf{C}'_X \setminus \{X\}} h(\mathbf{C}'_X) \Big/ \sum_{\mathbf{C}'_X} h(\mathbf{C}'_X)$.
11：**end for**

CollectMessage(**C**, **C**′)//**C**←**C**′

1：**for**（除 **C** 以外的每一个与 **C**′ 相邻的节点 **C**″
2：　　CollectMessage(**C**′, **C**″)；
3：**end for**
4：SendMessage(**C**′, **C**)；

DistributeMessage(**C**, **C**′)//**C**→**C**′

1：SendMessage(**C**, **C**′)；
2：**for**（除 **C** 以外的每一个与 **C**′ 相邻的节点 **C**″）
3：　　DistributeMessage（**C**′, **C**″）；
4：**end for**

SendMessage(**C**, **C**′)//**C**←**C**′

输入：**C**, **C**′——团树中的两个相邻的团；
作用：计算并储存从 **C** 到 **C**′ 的信息；

1：设 $\mathbf{C}'_1, \mathbf{C}'_2, \cdots, \mathbf{C}'_k$ 为 **C** 的除 **C**′ 以外的其它邻居；
2：对 $i=1, 2, \cdots, k$，$g_i \leftarrow$ RetrieveMessage($\mathbf{C}'_i$, **C**,)；
3：设 $f_1, f_2, \cdots, f_l$ 为初始化时储存在 **C** 处的函数，且设 $\mathbf{Z} = \mathbf{C} \backslash \mathbf{C}' \cup \mathbf{E}$；
4：计算如下函数：$\psi \leftarrow \sum_{\mathbf{Z}} \prod_{i=1}^{l} f_i \prod_{j=1}^{k} g_j$；
5：SaveMessage(**C**, **C**′, ψ).

图 5.7　团树传播算法 CTP

CTP 从 $\mathcal{T}$ 的叶节点开始，逐步朝枢纽节点方向传递信息. 在图 5.8 所示的团树中，如果选择 [TLR] 作为推理枢纽，那么信息收集阶段的信息传递情况如虚线箭头所示，其中，

$$\psi_1(T) = \sum_A P(A)P(T \mid A),$$

$$\psi_2(R) = \sum_X P(X \mid R),$$

$$\psi_3(L, B) = \sum_S P(S)P(L \mid S)P(B \mid S),$$

$$\psi_4(R, B) = P(D = y \mid R, B),$$

$$\psi_5(R, L) = \sum_B \psi_3(L,B) \cdot \psi_4(R,B).$$

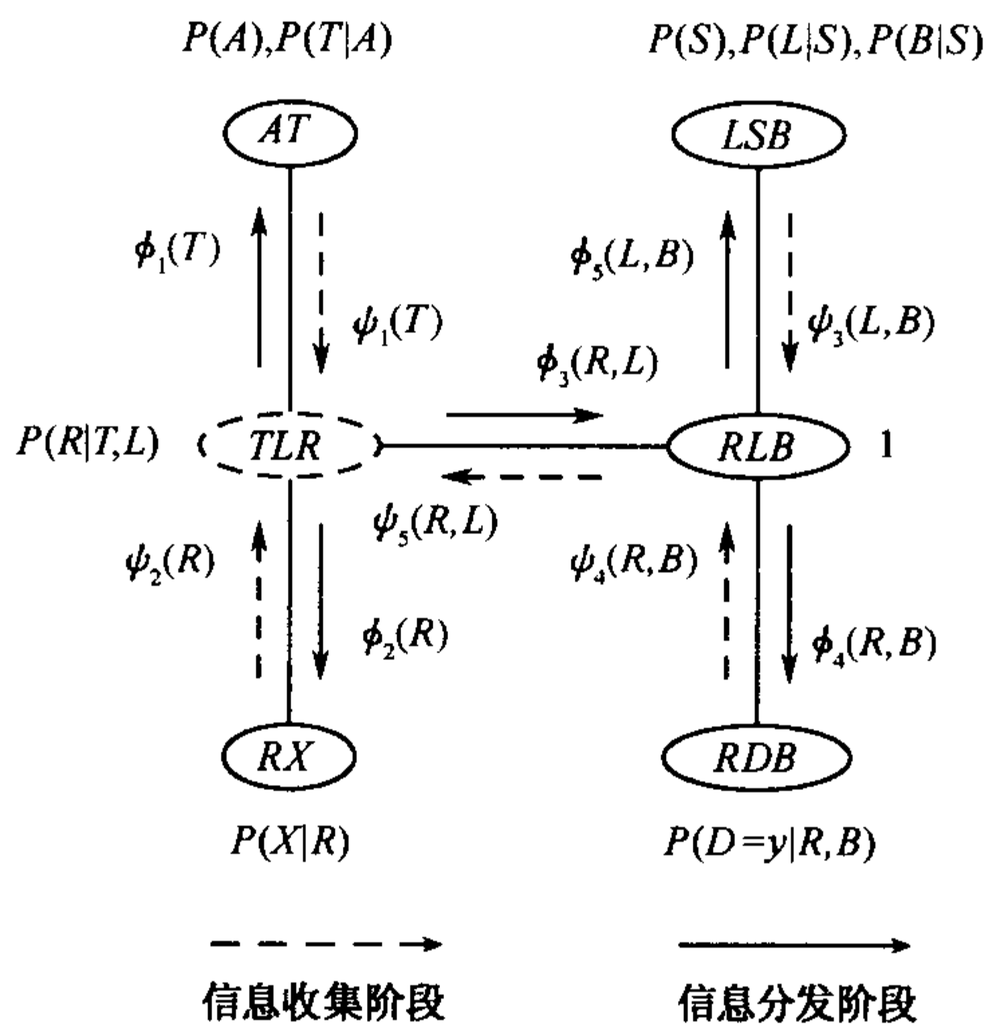


图 5.8　团树传播算法示例，这里 [TLR] 是枢纽节点

在信息分发阶段，即第 5 步，CTP 从枢纽节点 [TLR] 开始，逐步朝 $\mathcal{T}$ 的各个叶节点方向传递信息. 在图 5.8 中，信息分发阶段的信息传递情况如实线箭头所示，其中，

$$\phi_1(T) = \sum_{L,R} P(R \mid T,L)\psi_2(R)\ \psi_5(R, L)\ ,$$

$$\phi_2(R) = \sum_{L,T} P(R \mid T,L)\ \psi_1(T)\ \psi_5(R, L),$$

$$\phi_3(R, L) = \sum_T P(R \mid T,L)\ \psi_1(T)\ \psi_2(R),$$

$$\phi_4(R, B) = \sum_L \phi_3(R, L) \cdot \psi_3(L, B),$$

$$\phi_5(L, B) = \sum_R \phi_3(R, L) \cdot \psi_4(R,B) \cdot 1.$$

接下来的第 6 步称为**答案提取**（answer extraction）. 对每一个非证据变量 X，它首先找到一个包含 X 的团 $\mathbf{C}_X$，然后从 $\mathbf{C}_X$ 中“提取”后验分布 $P(X \mid \mathbf{E} = \mathbf{e})$. 在图 5.8 所示的例子中，$P(L \mid D = y)$ 可以从团 $[TLR]$ 中通过如下公式提取：

$$P(L \mid D = y) = \frac{\sum_{T, R} P(R \mid T, L)\psi_1(T)\psi_2(R)\psi_5(R,L)}{\sum_{T, R, L} P(R \mid T, L)\psi_1(T)\psi_2(R)\psi_5(R,L)}.$$

也可以从团 $[RLB]$ 中通过如下公式提取：

$$P(L \mid D = y) = \frac{\sum_{L, B} \psi_3(R, B) \cdot \psi_4(R, B) \cdot \phi_3(R, L) \cdot 1}{\sum_{L, R, B} \psi_3(L, B) \cdot \psi_4(R,B) \cdot \phi_3(R, L) \cdot 1}.$$

无论选择哪一个团，所得结果都一样.

把图 5.8 与图 5.3 相比较，可以看到 CTP 所计算、传递的信息是算法 CTP_0 的两倍. 前面曾经提过，CTP_0 本质上就是变量消元. 所以有时人们说，团树传播法可以用大约两倍于变量消元法计算一个变量的后验分布概率的时间，计算出每个非证据变量的后验概率分布.

CTP 计算了每一个非证据变量的后验概率分布. 如果只想计算部分变量的后验概率分布，只要对 CTP 的第 6 步进行适当的修改即可. 另外，如果一个变量的集合 $\mathbf{Z}$ 是 $\mathscr{T}$ 中某个团的子集合，那么也可以用 CTP 来计算 $\mathbf{Z}$ 的联合后验概率分布 $P(\mathbf{Z} \mid \mathbf{E} = \mathbf{e})$.

5.6 团树的构造

设 $\mathscr{N}$ 是一个贝叶斯网，本节考虑如何构造一棵覆盖 $\mathscr{N}$ 的团树.

5.6.1 图消元与团树

第 4.3 节在讨论变量消元法的计算复杂度时引入了图消元的概念，并引入了子程序 $\mathrm{Elim}(\mathscr{G}, Z)$. 图 5.9 给出了一个利用图消元来构造团树的算法，称为 BuildCT. 它的基本思想是：从贝叶斯网 $\mathscr{N}$ 的端正图 $\mathscr{G}$ 出发，按一定顺序在 $\mathscr{G}$ 中进行消元；在消去变量 X 之前，先构造一个由 X 以及所有与 X 相邻的变量组成的团；在消元过程结束后，把所产生的团以适当方式进行组织，得到一棵覆盖 $\mathscr{N}$ 的团树.

BuildCT ($\mathscr{G}$, ρ)

输入：$\mathscr{G}$——贝叶斯网 $\mathscr{N}$ 的端正图；

ρ——消元顺序.

输出：一个覆盖 $\mathscr{N}$ 的团树；

1：从 ρ 中删去第一个变量 Z；

2：$\mathbf{S} \leftarrow \mathrm{nb}(Z)$（所有与 Z 相邻的变量的集合），$\mathbf{C} \leftarrow \{Z\} \cup \mathbf{S}$；

3：**if**（**C** 包含 $\mathscr{G}$ 中所有节点）

　　return 以 **C** 为唯一节点的团树；

4：**else**

　　$\mathscr{G} \leftarrow$ Elim ($\mathscr{G}$, Z)；

　　$\mathscr{T} \leftarrow$ BuildCT ($\mathscr{G}$, ρ)；

5：**end if**

6：　　在 $\mathscr{T}$ 中找到一个包含 **S** 的团 $\mathbf{C}'$（后面我们会证明这样的团一定存在），将 **C** 与 $\mathbf{C}'$ 相连；

7：**return** $\mathscr{T}$.

图 5.9　团树构造的图消元算法

以图 5.1(c) 中所示的贝叶斯网 $\mathscr{N}$ 为例来逐步解释 BuildCT. $\mathscr{N}$ 的端正图 $\mathscr{G}$ 如图 5.10(a)所示，它是算法 BuildCT 的一个输入. 设另一个输入 ρ 是消元顺序：$\langle A, X, D, S, B, R, T, L \rangle$. BuildCT 首先从 ρ 中删除节点 A. 第 2 步 BuildCT构造两个集合，$\mathbf{C}_1 = \{A, T\}$，$\mathbf{S}_1 = \{T\}$. $\mathbf{C}_1$ 是消去 A 这一步所生成的团，而 $\mathbf{S}_1$ 把 A 和其它变量分隔开来，因此称为**分隔集**（separator）. 接下来，因为 $\mathbf{C}_1$ 不包含 $\mathscr{G}$ 中所有节点，所以算法跳过第 3 步而执行第 4 步，即在图 5.10(a) 中消去变量 A 并递归调用 BuildCT，其中消元顺序为 $\langle X, D, S, B, R, T, L \rangle$. 之后所产生的团和分隔集依次分别为：$\mathbf{C}_2 = \{R, X\}$，$\mathbf{S}_2 = \{R\}$；$\mathbf{C}_3 = \{R, B, D\}$，$\mathbf{S}_3 = \{R, B\}$；$\mathbf{C}_4 = \{L, S, B\}$，$\mathbf{S}_4 = \{L, B\}$；$\mathbf{C}_5 = \{R, L, B\}$，$\mathbf{S}_5 = \{L, R\}$；$\mathbf{C}_6 = \{R, T, L\}$，$\mathbf{S}_6 = \{T, L\}$. 在消去 R 时，所产生的团 $\mathbf{C}_6$ 包含了此时图 $\mathscr{G}$ 中的所有变量，于是 BuildCT 不再消去剩下的变量 T 和 L，而是逐层返回递归调用，开始构造团树（第 5 步）.

一开始团树 T 只有一个节点，即团 $\mathbf{C}_6$（见图 5.10(g)）. 在第一次执行 BuildCT 的第 5 步时，$\mathbf{C}_5$ 被加进 $\mathscr{T}$ 中：这时由于 $\mathbf{S}_5 \subseteq \mathbf{C}_6$，所以 $\mathbf{C}_5$ 被连到 $\mathbf{C}_6$（见图 5.10(h)）. 第二个被加进 $\mathscr{T}$ 的团是 $\mathbf{C}_4$：由于 $\mathbf{S}_4 \subseteq \mathbf{C}_5$，$\mathbf{C}_4$ 被连到 $\mathbf{C}_5$（见图 5.10(i)）；第三个被加进 $\mathscr{T}$ 的团是 $\mathbf{C}_3$：由于 $\mathbf{S}_3 \subseteq \mathbf{C}_5$，$\mathbf{C}_3$ 被连到 $\mathbf{C}_5$（见图 5.10(j)）；第四个被加进 $\mathscr{T}$ 的团是 $\mathbf{C}_2$：由于 $\mathbf{S}_2$ 同时是 $\mathbf{C}_6$、$\mathbf{C}_5$ 和 $\mathbf{C}_3$ 的子集，所以 $\mathbf{C}_2$ 可以与它们之中的任何一个相连. 假设 $\mathbf{C}_2$ 被连到 $\mathbf{C}_6$（见图 5.10(k)）. 最后一个被加进 $\mathscr{T}$ 的团是 $\mathbf{C}_1$：由于 $\mathbf{S}_1 \subseteq \mathbf{C}_6$，$\mathbf{C}_1$ 被连到 $\mathbf{C}_6$. BuildCT 得到的最后结果如图 5.10(l)所示，它与图 5.1(b)完全一样. 不难验证，它是一棵覆盖如图 5.1(c)所示的贝叶

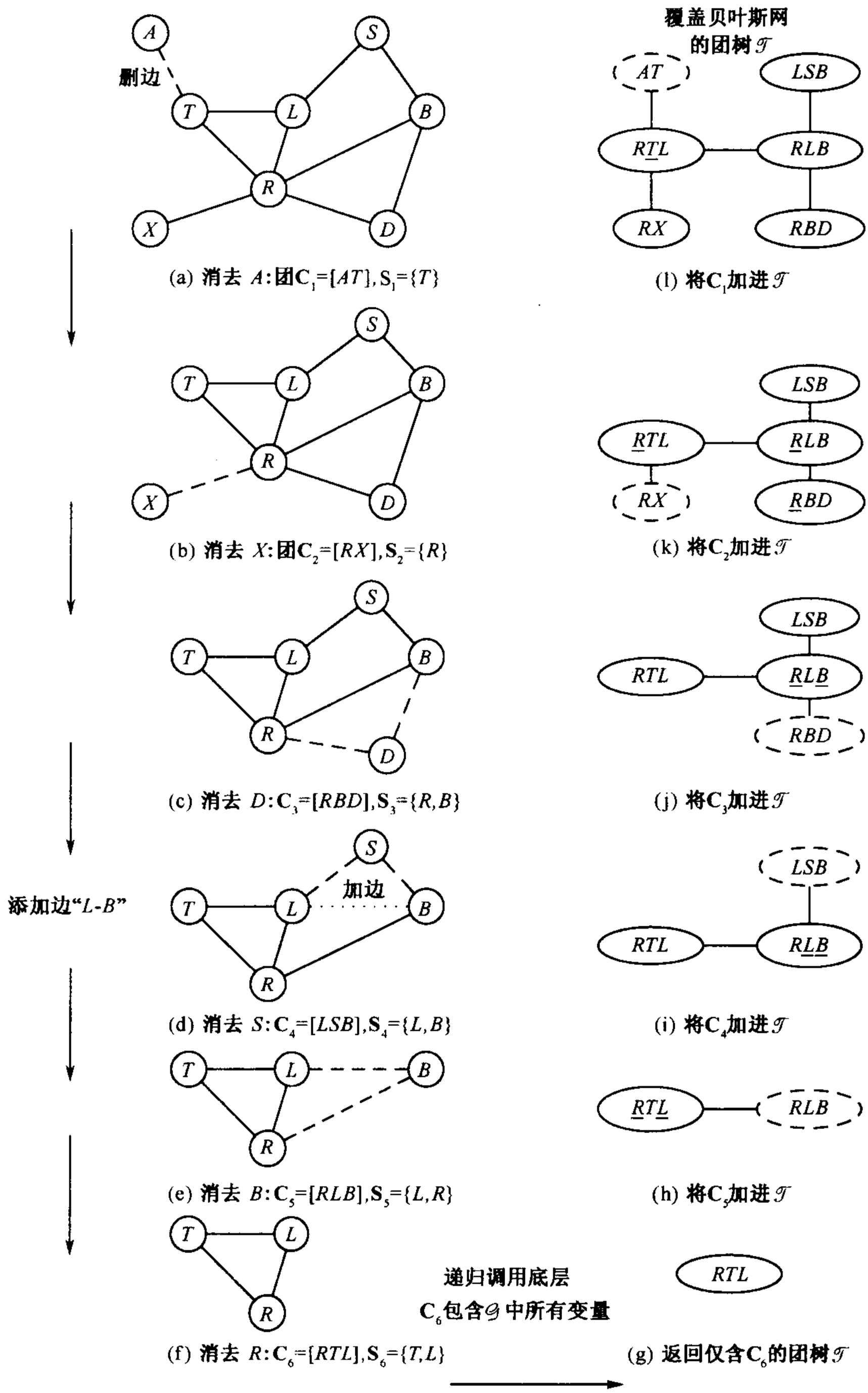

覆盖贝叶斯网
的团树$\mathcal{T}$
删边
添加边"L-B"
加边
递归调用底层
$\mathbf{C}_6$包含$\mathcal{G}$中所有变量
(a) 消去 A:团$\mathbf{C}_1$=[AT],$\mathbf{S}_1$={T}
(b) 消去 X:团$\mathbf{C}_2$=[RX],$\mathbf{S}_2$={R}
(c) 消去 D:$\mathbf{C}_3$=[RBD],$\mathbf{S}_3$={R,B}
(d) 消去 S:$\mathbf{C}_4$=[LSB],$\mathbf{S}_4$={L,B}
(e) 消去 B:$\mathbf{C}_5$=[RLB],$\mathbf{S}_5$={L,R}
(f) 消去 R:$\mathbf{C}_6$=[RTL],$\mathbf{S}_6$={T,L}
(g) 返回仅含$\mathbf{C}_6$的团树$\mathcal{T}$
(h) 将$\mathbf{C}_5$加进$\mathcal{T}$
(i) 将$\mathbf{C}_4$加进$\mathcal{T}$
(j) 将$\mathbf{C}_3$加进$\mathcal{T}$
(k) 将$\mathbf{C}_2$加进$\mathcal{T}$
(l) 将$\mathbf{C}_1$加进$\mathcal{T}$

图 5.10　图树构造示例

斯网 $\mathcal{N}$ 的团树.

5.6.2　图消元构造团树算法的正确性

设 **A** 是某无向图 $\mathcal{G}$ 中的一个子图. 如果 **A** 中每对节点都相邻，则称 **A** 是一个团 (clique). 有时候也称 **A** 的节点集合为团. 为了有别于团树中的团，把这里定义的团称为 ***g*-团**，而把团树中的团称为 ***t*-团**.

命题 5.2　设 **A** 是某无向图 $\mathcal{G}$ 中的一个 g-团，$\mathcal{T}$ 是 BuildCT 按某个消元顺序在 $\mathcal{G}$ 中进行消元而构造出的树. 那么，$\mathcal{T}$ 中一定存在一个 t-团 **C**，使得 $\mathbf{A} \subseteq \mathbf{C}$.

证明　将 $\mathcal{G}$ 中所有变量按顺序 ρ 排列. 假设 $A_1 \in \mathbf{A}$ 是 **A** 中排列最靠前的节点. 下面分两种情况证明命题.

(1) A_1 未被消去，即 $\mathrm{Elim}(\mathcal{G}, A_1)$ 未被调用时：由于 BuildCT 第 3 步产生的 t-团 **C** 包含当时 $\mathcal{G}$ 中所有变量，而当时所有 **A** 中的节点都未被消去，都在 $\mathcal{G}$ 中，所以 $\mathbf{A} \subseteq \mathbf{C}$.

(2) A_1 被消去并产生 t-团 **C** 时：由于 **A** 中所有其它节点都与 A_1 相邻，所以它们和 A_1 都在 **C** 中，即 $\mathbf{A} \subseteq \mathbf{C}$. 命题得证. □

命题 5.3　设 Z 是某无向图 $\mathcal{G}$ 中的一个节点，**S** 是所有与 Z 相邻的节点的集合. 设 $\mathcal{G}' = \mathrm{Elim}(\mathcal{G}, Z)$，而 $\mathcal{T}'$ 是 BuildCT 按某个消元顺序在 $\mathcal{G}'$ 中进行消元而构造出的树. 那么，在 $\mathcal{T}'$ 中一定有一个 t-团 $\mathbf{C}'$，使得 $\mathbf{S} \subseteq \mathbf{C}'$.

证明　在 $\mathcal{G}'$ 中，**S** 是一个 g-团. 从而根据命题 5.2，本命题得证. □

命题 5.4　设 $\mathcal{T}$ 是 BuildCT 按某个消元顺序在一个无向图中进行消元而构造出的树. 那么，$\mathcal{T}$ 是变量连通的，因而是团树.

证明　对无向图 $\mathcal{G}$ 的节点个数进行归纳. 当 $\mathcal{G}$ 只有一个节点时，命题显然成立. 假设命题在 $\mathcal{G}$ 有 k 个节点时成立，下面证明在 $\mathcal{G}$ 有 $k+1$ 个节点时命题也成立.

设 ρ 为消元顺序，Z 为 ρ 中第一个变量，$\mathbf{S} = \mathrm{nb}(Z)$，$\mathbf{C} = \{Z\} \cup \mathbf{S}$. 下面讨论两种情况：

(1) 当 **C** 包含 $\mathcal{G}$ 中所有节点时，BuildCT 返回一个只有一个节点 **C** 的团树，命题成立.

(2) 当 **C** 不能包含 $\mathcal{G}$ 中所有节点时，算法第 4 步首先消去 Z，得到一个节点个数为 k 的无向图 $\mathcal{G}'$，然后对 $\mathcal{G}'$ 调用 BuildCT 生成树 $\mathcal{T}'$. 根据归纳假设，$\mathcal{T}'$ 为一个 团树. 再根据命题 5.3，算法第 5 步可以在 $\mathcal{T}'$ 中找到一个包含 **S** 的 t-团 $\mathbf{C}'$. BuildCT 将 **C** 加入 $\mathcal{T}'$ 中，并将其与 $\mathbf{C}'$ 相连，得到团树 $\mathcal{T}$.

下面证明 $\mathcal{T}$ 也是变量连通的. 考虑 $\mathcal{T}$ 中所有包含某一节点 X 的 t-团所形成的子图 $\mathcal{T}_X$. 若 $X = Z$，则 $\mathcal{T}_X$ 中只有一个 t-团，即 **C**，所以它是连通的. 若 $X \notin \mathbf{C}$，则 $\mathcal{T}_X$ 也是 $\mathcal{T}'$ 中所有包含 X 的 t-团所形成的子图. 因为 $\mathcal{T}'$ 是一个团树，所以 $\mathcal{T}_X$ 是

连通的. 最后看 $X \neq Z$, 且 $X \in \mathbf{C}$ 的情况. 这时 $X \in \mathbf{S}$, 从而 $X \in \mathbf{C}'$. 设 $\mathcal{T}'_X$ 是 $\mathcal{T}'$ 中所有包含 X 的 t- 团所形成的子图. 因为 $\mathcal{T}'$ 是一个团树，所以 $\mathcal{T}'_X$ 是连通的. 由于 $X \in \mathbf{C}'$, $\mathbf{C}'$ 一定在 $\mathcal{T}'_X$ 中. 而 $\mathcal{T}_X$ 仅比 $\mathcal{T}'_X$ 多一个节点，即 $\mathbf{C}$. 又由于在 $\mathcal{T}$ 中 $\mathbf{C}'$ 和 $\mathbf{C}$ 相邻，所以 $\mathcal{T}_X$ 也是连通的. 从而 $\mathcal{T}$ 是变量连通的. 命题得证.

□

定理 5.2 设 $\mathcal{N}$ 是一贝叶斯网，$\mathcal{G}$ 是 $\mathcal{N}$ 的端正图，ρ 是 $\mathcal{N}$ 中所有变量的一个排序. 那么 BuildCT($\mathcal{G}$, ρ) 返回一棵覆盖 $\mathcal{N}$ 的团树.

证明 命题 5.3 意味着 BuildCT($\mathcal{G}$, ρ) 会运行成功（第 5 步一定能成功执行)，命题 5.4 意味着它的结果一定是一棵团树，记之为 $\mathcal{T}$. 对 $\mathcal{N}$ 中任一变量 X, $\{X\} \cup \pi(X)$ 是 $\mathcal{G}$ 中的一个 g- 团. 根据命题 5.2, $\mathcal{T}$ 中一定 有一个 t- 团 $\mathbf{C}$ 使得 $X \in \mathbf{C}$, 且 $\pi(X) \in \mathbf{C}$. 这就是说，$\mathcal{T}$ 是覆盖 $\mathcal{N}$ 的团树. 所以，BuildCT($\mathcal{G}$, ρ) 所返回的是一棵覆盖 $\mathcal{N}$ 的团树. 定理得证. □

5.6.3 极小团树

如果一棵团树中任意两相邻团都不互相包含，那么我们称之为**极小团树**. BuildCT 所构造的团树不一定是极小的. 例如，它在图 5.11(a)所示的无向图中，按 $\langle E, D, C, B, A\rangle$ 的顺序消元，就可能得到如图 5.11(b)所示的非极小团树. 图 5.11(b)所示的团树之所以不是极小的，是因为团 $[BC]$ 是其邻居 $[BCD]$ 的子集，而且 $[BCD]$ 是其邻居 $[BCDE]$ 的子集. 图 5.11(c)所示的团树也不是极小的，而图 5.11(d)所示的团树是极小的.

一棵非极小团树可以很容易被变为极小团树. 只需逐个考虑树中的团 $\mathbf{C}$, 如果有与 $\mathbf{C}$ 相邻的团 $\mathbf{C}'$ 使得 $\mathbf{C} \subset \mathbf{C}'$, 则从树中除去 $\mathbf{C}$, 并把 $\mathbf{C}$ 的除 $\mathbf{C}'$ 之外的其它邻居直接与 $\mathbf{C}'$ 相连. 这个过程称为**团树极小化**. 不难看出，所得到的新树仍保持变量连通性，因而是一个团树. 在图 5.11(b)所示的非极小团树中，由于团

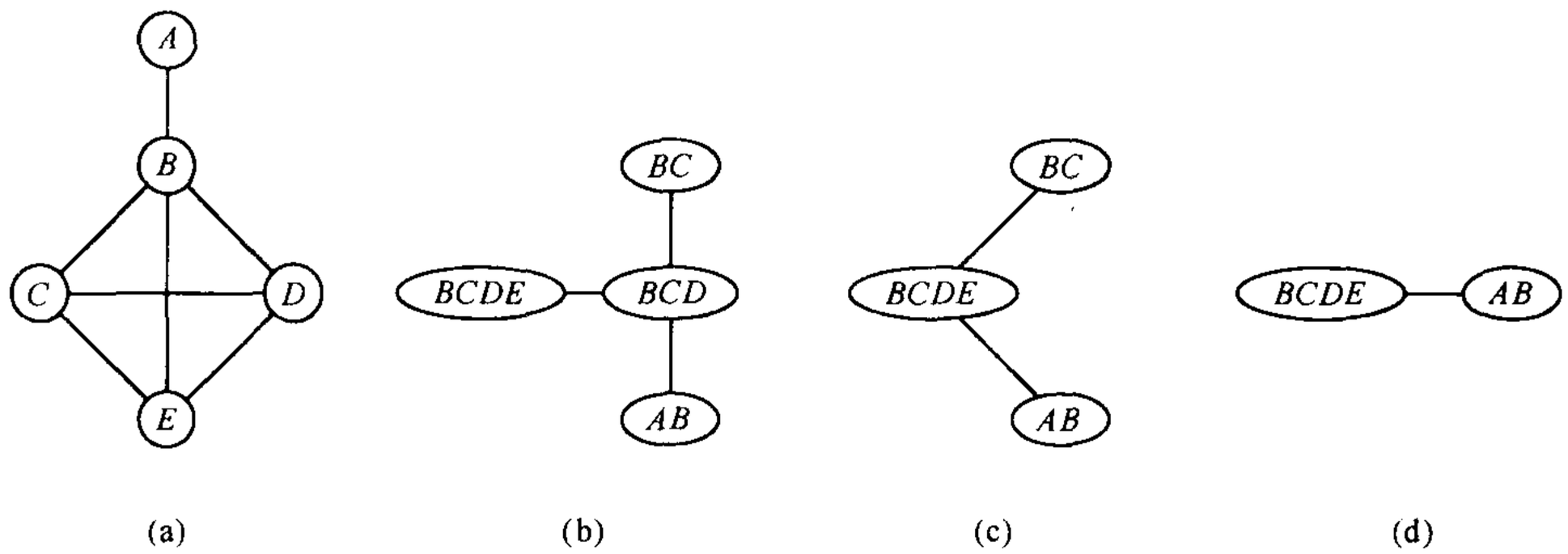


图 5.11 团树与极小团树

[*BCD*] 是 [*BCDE*] 的子集，所以把它除去，并将 [*BC*] 和 [*AB*] 连接到 [*BCDE*]，结果如图 5.11(c)所示；接着由于 [*BC*] 是 [*BCDE*] 的子集，所以再把它除去，最后得到如图 5.11(d)所示的极小团树.

值得注意的是，如果一棵团树 $\mathscr{T}$ 覆盖一个贝叶斯网 $\mathscr{N}$，将 $\mathscr{T}$ 极小化后，它仍然覆盖 $\mathscr{N}$. 从现在起规定，BuildCT 在输出它所构造的团树之前要先将其极小化.

5.6.4　*t*-团与 *g*-团

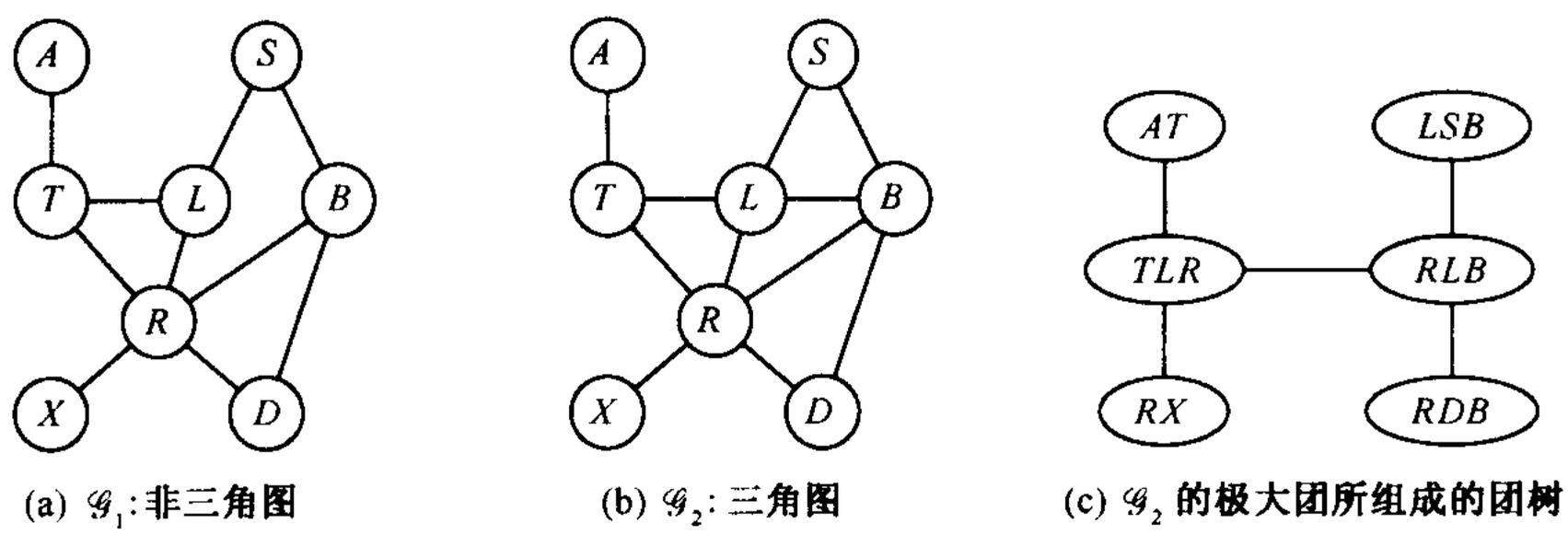
下面讨论 *t*-团与 *g*-团之间的联系. 我们将会看到，BuildCT 所构造的 *t*-团其实是某个无向图中的 *g*-团. 这就是团树之所以被称为是团树，以及其中的节点被称为是团的原因.

无向图中的一个**圈**（cycle）是一个节点序列，其中节点各异、且每对相邻节点以及首尾节点之间都有一条边直接相连. 连接圈中两个不相邻节点的边称为**弦**（chord). 如果一个无向图中的每个包含 4 个或 4 个以上节点的圈都有一条弦，那么该图称为**弦图**（chordal graph)，或**三角图**（triangulated graph). 图 5.12(a)所示的 $\mathscr{G}_1$ 不是一个三角图，因为圈“S-L-R-B”没有弦. 图 5.12(b)所示的 $\mathscr{G}_2$ 是一个三角图.

无向图中一些节点的集合称为**极大团**（maximal clique)，如果它是一个 *g*-团，而且它的任何**超集**（super set）都不是 *g*-团. 三角图的一个重要特性是，它的所有极大团可以组织成一个团树. 例如，图 5.12(b) 中的三角图有 6 个极大团：[*AT*]、[*TLR*]、[*RX*]、[*LSB*]、[*RLB*] 和[*RDB*]. 这些团可以组织成如图 5.12(c)所示的团树.

一个非三角图可以通过添加边而变成三角图，这个过程叫**三角化**（triangulation). 例如，图 5.12(a)中的 $\mathscr{G}_1$ 不是三角图，但如果在其中添加一条边“*L-B*”，就得到图 5.12(b)中的三角图 $\mathscr{G}_2$.

考虑按照某个顺序 ρ 在一个无向图 $\mathscr{G}$ 中进行消元的过程. 在消去某个节点 Z



(a) $\mathscr{G}_1$：非三角图　　(b) $\mathscr{G}_2$：三角图　　(c) $\mathscr{G}_2$ 的极大团所组成的团树

图 5.12　三角图与团树

时，可能会需要添加一些边使得所有与 Z 相邻的节点两两相连. 如果把消元过程中所有这些添加的边都记录下来，再把它们添加到消元之前的无向图 $\mathscr{G}$ 中，则可得到另一个无向图，记之为 $\mathscr{G}_+$. 可以证明，$\mathscr{G}_+$ 是一个三角图. 作为例子，考虑在图 5.12(a)所示的 $\mathscr{G}_1$ 中按顺序 $\langle A, X, D, S, B, R, T, L \rangle$ 进行消元的过程. 在消去 S 时，需要添加边 "L-B". 如果将这条边添加到 $\mathscr{G}_1$ 中，就得到三角图 $\mathscr{G}_2$.

设 $\mathscr{T}$ 是 BuildCT($\mathscr{G}$, ρ) 所构造的团树. 按照 5.6.3 小节最后的假设，$\mathscr{T}$ 已经被极小化. 可以证明：**$\mathscr{T}$ 中的 t-团就是无向图 $\mathscr{G}_+$ 中的极大 g-团**. 这就是本小节所要指出的重要事实. 例如，考虑如图 5.10 所示的图消元构造团树的过程，它所构造的团树中的 t-团，正好是 $\mathscr{G}_2$ 中的极大团.

本节的讨论自然会让读者联想到如下的构造覆盖贝叶斯网 $\mathscr{N}$ 的团树的方法：

(1) 计算 $\mathscr{N}$ 的端正图 $\mathscr{G}$；

(2) 将 $\mathscr{G}$ 三角化，得到三角图 $\mathscr{G}_T$；

(3) 找到 $\mathscr{G}_T$ 中的所有极大团；

(4) 将这些团组织成一棵团树 $\mathscr{T}$.

在以前的文献中，这是构造团树的标准方法. 我们之所以选择介绍 BuildCT，是因为它无论在概念上还是在实现上，都比上述的标准方法更为简单.

5.7 文献介绍

团树传播法有两个版本：一是 Hugin 方法（Lauritzen and Spiegelhalter, 1988; Jensen et al., 1990），二是本书所介绍的 Shafer-Shenoy 方法（Shafer and Shenoy, 1990). 这两个版本基本等价，但后者在概念上更为简单. 从变量消元和计算共享的角度理解团树传播是本书的特点之一. Madsen 和 Jensen (1998) 注意到算法实现的具体方法有时对团树传播的效率有重要影响，因此提出了惰性传播（lazy propagation）法. Zhang (1998a) 对团树传播和变量消元进行了比较.

三角图的概念最早是在研究解线性方程组的高斯消元法时提出的（Rose, 1970; Tarjan et al., 1976），后来人们发现它对关系数据库理论也至关重要 (Beeri et al., 1983). 在数据库理论中，团树也称为"join tree" (Maier, 1983). 读者若想进一步了解三角图与团树之间的关系，可以参阅有关文献（Cowell et al., 1999（第 4 章）; Golumbic, 1980; Tarjan and Yannakakis, 1984; Zhang, 1993).

第6章 近似推理

前两章介绍的变量消元法和团树传播法都是精确推理算法. 当网络节点众多并且连接稠密时，它们的计算复杂度高，因而精确推理算法不适用. 这时候通常有两个选择：一是利用概率分布中的局部结构，如因果机制独立和环境独立（见第2章），来进一步降低精确推理的复杂度；另一个是降低对精度的要求，考虑近似推理算法，以求在限定的时间内得到一个近似解. 本章讨论近似推理算法，首先介绍随机抽样法和变分法，然后简单介绍一些基于模型简化和基于搜索的方法.

6.1 随机抽样算法

随机抽样算法（random sampling algorithms），又称为随机仿真（random simulation）或蒙特卡洛算法（Monte Carlo algorithms），是一类广泛应用于数值积分和统计物理的近似计算方法. 它的基本思想是从某个概率分布随机抽样，生成一组样本，然后从样本出发近似估计要计算的量.

随机抽样算法可以分为两大类，即**重要性抽样**（importance sampling）算法和**马尔可夫链蒙特卡洛**（Markov chain Monte Carlo）算法，后者又简称为MCMC算法. 它们的主要区别在于前者产生的样本之间相互独立，而后者产生的样本却相互关联. 下面分别介绍它们在贝叶斯网推理中的应用.

6.1.1 重要性抽样法

1. 重要性抽样法基本理论

设 $f(\mathbf{X})$ 是一组变量 $\mathbf{X}$ 在其定义域 $\Omega_{\mathbf{X}} \subset R^n$ 上的可积函数. 考虑计算积分

$$I = \int_{\Omega_{\mathbf{X}}} f(\mathbf{X}) \mathrm{d}\mathbf{X}. \tag{6.1}$$

为了近似计算这一积分，重要性抽样法将上式改写为如下形式：

$$I = \int_{\Omega_{\mathbf{X}}} \frac{f(\mathbf{X})}{p(\mathbf{X})} p(\mathbf{X})\ \mathrm{d}\mathbf{X}, \tag{6.2}$$

这里，$\mathbf{X}$ 被看成是一组随机变量，$p(\mathbf{X})$ 是 $\mathbf{X}$ 的一个联合概率分布，称为**重要性分布**（importance distribution），它满足以下条件：对 $\mathbf{X}$ 的任一取值 $\mathbf{x}$，如果 $f(\mathbf{X}=\mathbf{x}) \neq 0$，那么 $p(\mathbf{X}=\mathbf{x}) \neq 0$.

接下来，重要性抽样法从 $p(\mathbf{X})$ 独立地抽取 m 个样本 $\mathbf{D}_1,\mathbf{D}_2,\cdots,\mathbf{D}_m$，并基于这些样本来对积分 I 进行估计：

$$\hat{I}_m = \frac{1}{m}\sum_{i=1}^{m}\frac{f(\mathbf{D}_i)}{p(\mathbf{D}_i)}. \tag{6.3}$$

可以证明，$\hat{I}_m$ 是 I 的一个无偏估计，且根据强大数定律，当样本量 m 趋于无穷时，$\hat{I}_m$ 几乎必然收敛于 I.

重要性抽样法的性能主要从两个方面来衡量：一个是算法复杂度，另一个是近似解的精度. 因此，人们用计算 $\hat{I}_m$ 所需的时间 t 和 $\hat{I}_m$ 的方差 $\mathrm{var}(\hat{I}_m)$ 之积 $t\cdot\mathrm{var}(\hat{I}_m)$ 来度量重要性抽样法的**效率**（efficiency）：$t\cdot\mathrm{var}(\hat{I}_m)$ 越小，算法的效率越高，收敛速度也就越快，从而获得高精度近似所需的样本量不大. 这里，方差可用下式计算：

$$\mathrm{var}(\hat{I}_m) = \frac{1}{m}\left[\int_{\Omega_{\mathbf{X}}}\frac{f^2(\mathbf{X})}{p(\mathbf{X})}\mathrm{d}\mathbf{X} - I^2\right]. \tag{6.4}$$

重要性分布的选择是提高算法效率的关键. 由于重要性分布的选择对时间复杂度 t 的影响不大，因此为了提高算法的效率，应该选用使得方差 $\mathrm{var}(\hat{I}_m)$ 尽可能小的重要性分布. 根据式（6.4），若被积函数 $f(\mathbf{X})>0$，则**最优重要性分布**为 $p^*(\mathbf{X}) = f(\mathbf{X})/I$. 此时 $\mathrm{var}(\hat{I}_m)=0$，样本被集中在 $f(\mathbf{X})$ 值较大的“重要”区域. 由于 I 本身是未知的，在实际中很少能够从 $p^*(\mathbf{X})$ 抽样，只能寻找与 $p^*(\mathbf{X})$ 尽量接近的分布. 重要性分布与最优分布 $p^*(\mathbf{X})$ 越接近，方差 $\mathrm{var}(\hat{I}_m)$ 就越小.

上面的这些讨论和结果是针对连续变量的，它们可以自然地推广到离散变量的情况，此时，积分运算相应地被求和运算替换. 若读者想进一步了解重要性抽样法，可以参见文献（Rubinstein，1981）.

2. 重要性抽样法与概率推理

考虑一个贝叶斯网 $\mathcal{N}$，用 $\mathbf{X}$ 记其中所有变量的集合，$P(\mathbf{X})$ 记 $\mathcal{N}$ 所表示的联合概率分布. 设观测到证据 $\mathbf{E}=\mathbf{e}$. 下面将讨论如何近似计算一组查询变量 $\mathbf{Q}$ 取某值 $\mathbf{q}$ 的后验概率 $P(\mathbf{Q}=\mathbf{q}\mid\mathbf{E}=\mathbf{e})$.

设 $\mathbf{W}$ 是一些变量的集合，$\mathbf{Y}$ 是 $\mathbf{W}$ 的一个子集合，$\mathbf{Z}=\mathbf{W}\setminus\mathbf{Y}$，并设 $\mathbf{y}$ 为 $\mathbf{Y}$ 的一个取值. 定义函数

$$\chi_{\mathbf{Y}=\mathbf{y}}(\mathbf{W}) = \chi_{\mathbf{Y}=\mathbf{y}}(\mathbf{Y},\mathbf{Z}) = \begin{cases}1, \text{若 } \mathbf{Y}=\mathbf{y}\\ 0, \text{若否}\end{cases}. \tag{6.5}$$

按条件概率的定义，有

$$P(\mathbf{Q}=\mathbf{q}\mid\mathbf{E}=\mathbf{e}) = \frac{P(\mathbf{Q}=\mathbf{q},\mathbf{E}=\mathbf{e})}{P(\mathbf{E}=\mathbf{e})}. \tag{6.6}$$

$P(\mathbf{Q}=\mathbf{q},\mathbf{E}=\mathbf{e})$ 和 $P(\mathbf{E}=\mathbf{e})$ 可以分别表示成如下期望形式：

$$P(\mathbf{Q}=\mathbf{q},\mathbf{E}=\mathbf{e})=\sum_{\mathbf{X}}\chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{X})\ \chi_{\mathbf{E}=\mathbf{e}}(\mathbf{X})P(\mathbf{X}), \tag{6.7}$$

$$P(\mathbf{E}=\mathbf{e})=\sum_{\mathbf{X}}\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{X})P(\mathbf{X}). \tag{6.8}$$

于是可以利用重要性抽样法来对它们进行近似. 下面将给出两种不同的具体方法.

对于近似的一般性质，有一点需要注意. 根据以上讨论，利用重要性抽样法获得的对 $P(\mathbf{Q}=\mathbf{q},\mathbf{E}=\mathbf{e})$ 和 $P(\mathbf{E}=\mathbf{e})$ 的估计是无偏的. 但是从它们出发，通过式（6.6）获得的对 $P(\mathbf{Q}=\mathbf{q}\mid\mathbf{E}=\mathbf{e})$ 的估计却不是无偏的. 所幸的是，偏倚基本上随着样本量的增加而减少，当样本量趋于无穷时，这个估计几乎必然收敛于真实值.

3. 逻辑抽样法

要用重要性抽样法解决式（6.7）和式（6.8）定义的问题，首先需要选择一个重要性分布. 一个很自然的想法就是选用联合分布 $P(\mathbf{X})$ 本身来作为重要性分布，这样就得到了**逻辑抽样**（logic sampling).

有向无圈图的一个**拓扑序**（topological sort）是图中所有节点的一个线性序，其中每个节点都在它的子节点之前出现.

逻辑抽样法首先从分布 $P(\mathbf{X})$ 中抽取样本. 注意到 $P(\mathbf{X})$ 可以分解为

$$P(\mathbf{X})=\prod_{X\in\mathbf{X}}P(X\mid\pi(X)),$$

因此可以按照贝叶斯网 $\mathcal{N}$ 的拓扑序对其中的变量逐个进行抽样：对待抽样变量 X，若它是根节点，则按分布 $P(X)$ 进行抽样；若是非根节点，则按分布是 $P(X\mid\pi(X)=r)$ 进行抽样，这里 $\pi(X)=r$ 是 X 的父节点的抽样结果，在对 X 抽样时是已知的. 这样的抽样过程称为**顺序抽样**（forward sampling).

顺序抽样过程需要从一些单变量概率分布随机抽样，这可以借助一个**随机数产生器**（random number generator）来实现. 例如，对于一个服从伯努利分布的二值随机变量 X，设 $\Omega_X=\{0,1\}$，$P(X=0)=p$，$P(X=1)=1-p$. 从 $P(X)$ 抽样可以这样进行：利用随机数产生器产生一个实数 $r\in[0,1]$，若 $r\in[0,p]$，则生成样本 0，否则生成样本 1. 如果随机数产生器产生的实数服从 $[0,1]$ 上的均匀分布，则这样生成的样本服从分布 $P(X)$. 上述过程可以自然地推广到 X 取多个值的情况.

例 6.1 对图 6.1 所示的贝叶斯网顺序抽样. 首先对根节点 C 抽样，抽样分布是 $P(C)$. 假设抽样结果为 $C=t$. 接下来分别对节点 R 和 S 抽样，因为它们的父节点取值已经确定为 $C=t$，所以抽样分布分别为 $P(R\mid C=t)$ 和 $P(S\mid C=t)$. 假设抽样结果分别为 $R=t$ 和 $S=f$. 最后对叶节点 W 抽样，抽样分布为

$P(W \mid R=t, S=f)$. 假设抽样结果为 $W=t$. 这样，最终生成的样本为 $\{C=t, R=t, S=f, W=t\}$. □

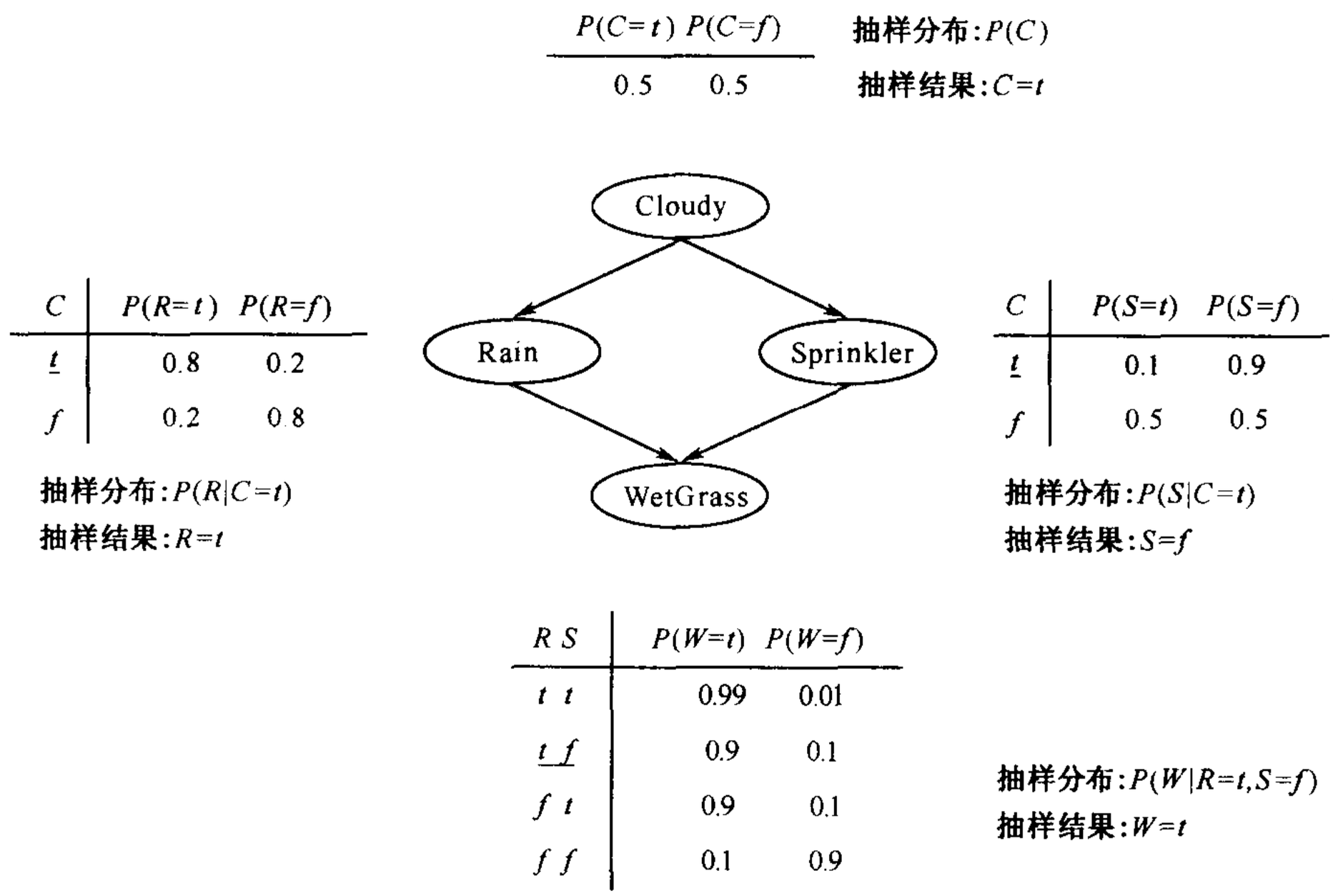


图 6.1 顺序抽样示例，所得样本 $\{C=t, R=t, S=f, W=t\}$

假设通过顺序抽样过程获得了 m 个独立样本 $\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m$，其中满足 $\mathbf{E}=\mathbf{e}$ 的有 $m_{\mathbf{e}}$ 个，而在这 $m_{\mathbf{e}}$ 个样本中，进一步满足 $\mathbf{Q}=\mathbf{q}$ 的有 $m_{\mathbf{q},\mathbf{e}}$ 个. 根据式 (6.3) 和式 (6.7)，有

$$
\begin{aligned}
P(\mathbf{Q}=\mathbf{q}, \mathbf{E}=\mathbf{e}) &\approx \frac{1}{m}\sum_{i=1}^{m}\frac{\chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{D}_i)\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{D}_i)P(\mathbf{D}_i)}{P(\mathbf{D}_i)} \\
&= \frac{1}{m}\sum_{i=1}^{m}\chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{D}_i)\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{D}_i) \\
&= \frac{m_{\mathbf{q},\mathbf{e}}}{m}.
\end{aligned}
$$

类似地，根据式 (6.3) 和式 (6.8) 可得

$$
P(\mathbf{E}=\mathbf{e}) \approx \frac{1}{m}\sum_{i=1}^{m}\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{D}_i) = \frac{m_{\mathbf{e}}}{m}.
$$

将上面两式代入式 (6.6)，得

$$
P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e}) \approx \frac{m_{\mathbf{q},\mathbf{e}}}{m_{\mathbf{e}}}, \tag{6.9}
$$

这就是通过逻辑抽样法获得的对后验概率的近似，它是在所有满足 $\mathbf{E}=\mathbf{e}$ 的样本

中，进一步满足 $\mathbf{Q}=\mathbf{q}$ 的样本所占的比例.

逻辑抽样法所产生与证据 $\mathbf{E}=\mathbf{e}$ 不一致的那些样本相当于被舍弃. 因此，逻辑抽样有时也称为**舍选抽样**（rejection sampling).

例 6.2　对如图 6.1 所示的贝叶斯网，用逻辑抽样近似计算 $P(R=t \mid S=t)$. 假设总样本数 $m=100$，其中有 75 个满足 $S=f$. 因为它们与观测证据不一致，所以被舍弃. 在另外 25 个满足 $S=t$ 的样本中，有 18 个满足 $R=f$. 因此得到的近似解为 $P(R=t \mid S=t) \approx 18/25=0.72$，而它的精确值为 0.7. □

逻辑抽样的优点是简单易行，缺点是当概率 $P(\mathbf{E}=\mathbf{e})$ 很小时，算法效率低，收敛速度慢. 事实上，问题的最优重要性分布是 $\frac{\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{X})P(\mathbf{X})}{P(\mathbf{E}=\mathbf{e})}$，而抽样使用的是分布 $P(\mathbf{X})$，两者差别显著：前者的概率质量集中在 $\mathbf{X}$ 的与 $\mathbf{E}=\mathbf{e}$ 一致的那些取值处，而后者在这些区域的概率值却很小. 随着 $P(\mathbf{E}=\mathbf{e})$ 的减小，所抽得的与 $\mathbf{E}=\mathbf{e}$ 一致的样本个数将会减少，因此大量样本被舍弃，造成了计算资源的浪费.

4. 似然加权法

似然加权（likelihood weighting）是重要性抽样法的另外一个特例，提出它的主要目的是要避免逻辑抽样因舍弃样本而造成的浪费. 图 6.2 给出了似然加权算法的伪代码，其中第 4～12 行是抽样过程. 它按拓扑序对每个变量 X 进行抽样：当 X 不是证据变量时，抽样方式与逻辑抽样法一样；而当 X 是证据变量时，则以 X 的观测值作为抽样结果. 这样保证了每一个样本都与证据 $\mathbf{E}=\mathbf{e}$ 一致，从而可以利用，不必舍弃.

例 6.3　对如图 6.1 所示的贝叶斯网，用似然加权算法计算 $P(R=t \mid S=t)$. 似然加权算法生成一个样本的过程如下：

(1) 对根节点 C，从 $P(C)$ 抽样，假设得到 $C=t$；

(2) 对节点 S，因为 $S \in \mathbf{E}$ 是证据变量，所以抽样结果被设为它的观测值 t；

(3) 对节点 R，抽样分布为 $P(R \mid C=t)$，假设得到 $R=t$；

(4) 最后对叶节点 W 抽样，抽样分布为 $P(W \mid R=t, S=t)$，假设得到 $W=t$. 最后产生的样本为 $\mathbf{D}=\{C=t, R=t, S=t, W=t\}$. □

设 $\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m$ 是通过上述过程抽得的 m 个样本. 下面讨论怎样基于它们对 $P(\mathbf{Q}=\mathbf{q}, \mathbf{E}=\mathbf{e})$ 和 $P(\mathbf{E}=\mathbf{e})$ 进行近似.

设 $\mathbf{Y}$ 是 $\mathbf{X}$ 的一个子集. 对任一 $\mathbf{Y}$ 的函数 $h(\mathbf{Y})$，用 $h(\mathbf{Y})|_{\mathbf{D}_i}$ 记当变量取 $\mathbf{D}_i$ 中的值时，这个函数的函数值. 对任一 $X \in \mathbf{X}$，$P(X \mid \pi(X))$ 是 $\mathbf{X}$ 中一些变量的函数. 于是，$P(X \mid \pi(X))|_{\mathbf{D}_i}$ 是当变量取 $\mathbf{D}_i$ 中的值时，这个函数的函数值.

用 $\mathbf{Z}$ 记所有非证据变量的集合，即 $\mathbf{Z}=\mathbf{X} \setminus \mathbf{E}$. 设 $P'(\mathbf{X})$ 是似然加权算法所使用的重要性分布. 不难看出，$P'(\mathbf{E})=\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{E})$，而

LikelihoodWeighting($\mathcal{N}$, m,**E**,**e**,**Q**,**q**)

输入:$\mathcal{N}$——一个贝叶斯网;m——样本量;
　　E——证据变量;**e**——证据变量的取值;
　　Q——查询变量;**q**——查询变量的取值.
输出:对 $P(\mathbf{Q}=\mathbf{q}|\mathbf{E}=\mathbf{e})$的近似;

```
1: ρ←N的一个拓扑序;
2: w_e←0;w_{q,e}←0;
3: for(i=1 to m)
4:     D_i←∅ ;
5:     for(ρ中的每一个变量X)
6:         if(X∈E)
7:             x←X的观测值;
8:         else
9:             x←从P(X|π(X))抽样的结果;
10:        end if
11:    end for
12:    D_i←D_i∪{X=x};
13:    w_i ← ∏_{X∈E} P(X | π(X)) |_{D_i} ;
14:    w_e←w_e+w_i;
15:    if(D与D=q一致)
16:        w_{q,e}←w_{q,e}+w_i;
17:    end if
18: end for
19: return w_{q,e}/w_e.
```

图 6.2　似然加权算法

$$P'(\mathbf{Z} \mid \mathbf{E}) = P(\mathbf{Z} \mid \mathbf{E}) = \prod_{X \in \mathbf{Z}} P(X \mid \pi(X)).$$

于是有

$$\begin{aligned} P'(\mathbf{X}) &= P'(\mathbf{Z},\mathbf{E}) \\ &= P'(\mathbf{E})\ P'(\mathbf{Z} \mid \mathbf{E}) \\ &= \chi_{\mathbf{E}=\mathbf{e}}(\mathbf{E}) \prod_{X \in \mathbf{Z}} P(X \mid \pi(X)). \end{aligned}$$

注意,$\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{X}) = \chi_{\mathbf{E}=\mathbf{e}}(\mathbf{E},\mathbf{Z}) = \chi_{\mathbf{E}=\mathbf{e}}(\mathbf{E})$. 于是有

$$P'(\mathbf{X}) = \chi_{\mathbf{E}=\mathbf{e}}(\mathbf{X}) \prod_{X \in \mathbf{Z}} P(X \mid \pi(X)).$$

对每个 $i=1,2,\cdots,m$,样本 $\mathbf{D}_i$ 与证据 $\mathbf{E}=\mathbf{e}$ 一致,因此对任一函数 $h(\mathbf{X})$,有 $\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{D}_i)h(\mathbf{D}_i) = h(\mathbf{X})\mid_{\mathbf{D}_i}$. 所以,

$$P'(\mathbf{X}) = \prod_{X \in \mathbf{Z}} P(X \mid \pi(X)) \mid_{\mathbf{D}_i}.$$

$$\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{D}_i)\, P(\mathbf{D}_i) = \prod_{X \in \mathbf{X}} P(X \mid \pi(X)) \mid_{\mathbf{D}_i}.$$

根据式（6.3）和式（6.7），可得

$$\begin{aligned} P(\mathbf{Q}=\mathbf{q}, \mathbf{E}=\mathbf{e}) &\approx \frac{1}{m}\sum_{i=1}^{m} \frac{\chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{D}_i)\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{D}_i)P(\mathbf{D}_i)}{P'(\mathbf{D}_i)} \\ &= \frac{1}{m}\sum_{i=1}^{m} \chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{D}_i) \frac{\prod_{X \in \mathbf{X}} P(X \mid \pi(X)) \mid_{\mathbf{D}_i}}{\prod_{X \in \mathbf{Z}} P(X \mid \pi(X)) \mid_{\mathbf{D}_i}} \\ &= \frac{1}{m}\sum_{i=1}^{m} \chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{D}_i) \prod_{X \in \mathbf{E}} P(X \mid \pi(X)) \mid_{\mathbf{D}_i} \\ &= \frac{1}{m}\sum_{i=1}^{m} \chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{D}_i) w(\mathbf{D}_i), \end{aligned} \tag{6.10}$$

其中 $w(\mathbf{D}_i) = \prod_{X \in \mathbf{E}} P(X \mid \pi(X)) \mid_{\mathbf{D}_i}$. 类似地，根据式（6.3）和式（6.8），可得

$$\begin{aligned} P(\mathbf{E}=\mathbf{e}) &\approx \frac{1}{m}\sum_{i=1}^{m} \frac{\chi_{\mathbf{E}=\mathbf{e}}(\mathbf{D}_i)P(\mathbf{D}_i)}{P'(\mathbf{D}_i)} \\ &= \frac{1}{m}\sum_{i=1}^{m} w(\mathbf{D}_i). \end{aligned} \tag{6.11}$$

将式（6.10）和式（6.11）代入式（6.6），得

$$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e}) \approx \frac{\sum_{i=1}^{m} \chi_{\mathbf{Q}=\mathbf{q}}(\mathbf{D}_i) w(\mathbf{D}_i)}{\sum_{i=1}^{m} w(\mathbf{D}_i)}. \tag{6.12}$$

与逻辑抽样相比，似然加权算法相当于为每个样本 $\mathbf{D}_i$ 都赋予一个权重 $w(\mathbf{D}_i)$. 设 $\mathbf{z}_i$ 为 $\mathbf{D}_i$ 中 $\mathbf{Z}$ 的取值，则 $\mathbf{D}_i$ 可以写成 $\mathbf{D}_i = (\mathbf{Z}=\mathbf{z}_i, \mathbf{E}=\mathbf{e})$. 当所有证据变量都是叶节点时，权重 $w(\mathbf{D}_i)$ 是 $P(\mathbf{E}=\mathbf{e} \mid \mathbf{Z}=\mathbf{z}_i)$，即给定 $\mathbf{E}=\mathbf{e}$ 时 $\mathbf{Z}=\mathbf{z}_i$ 的似然度. 这就是似然加权算法得名的原因.

似然加权的优点是每个样本都被利用，效率比逻辑抽样有很大提高. 当所有证据变量都位于网络顶端时，重要性分布 $P'(\mathbf{X})$ 正好就是最优分布，似然加权的效率达到最优. 在其它情况下，重要性分布与最优分布可能差别显著，尤其是当概率 $P(\mathbf{E}=\mathbf{e})$ 很小时. 这时，算法的收敛会变慢，即要获得高精度的近似所需的样本量会增加.

6.1.2 MCMC 抽样

在重要性抽样算法中，不同样本之间相互独立. 下面介绍 MCMC 抽样算法，

其中不同样本之间不是相互独立的.

图 6.3 所示的是一个简单的 MCMC 算法，称为**吉布斯抽样**（Gibbs sampling）. 它首先随机生成一个与证据 $\mathbf{E}=\mathbf{e}$ 相一致的样本 $\mathbf{D}_1$ 作为起始样本，此后每一步都从当前样本出发产生下一个样本. 设当前在 $i-1$ 步，为了从 $\mathbf{D}_{i-1}$ 出发得到 $\mathbf{D}_i$，算法首先设 $\mathbf{D}_i=\mathbf{D}_{i-1}$，然后按照某个顺序对非证据变量逐个进行抽样，改变 $\mathbf{D}_i$ 中变量的取值. 设 Z 是下一个待抽样变量，$\mathbf{Y}$ 是 Z 的马尔可夫边界上的变量的集合，$\mathbf{y}_i$ 是 $\mathbf{Y}$ 在 $\mathbf{D}_i$ 中的当前取值. 算法根据分布 $P(Z\mid\mathbf{Y}=\mathbf{y}_i)$ 对 Z 进行抽样，并用抽样结果替代 $\mathbf{D}_i$ 中 Z 的当前取值.

GibbsSampling ($\mathcal{N}$, m, $\mathbf{E}$, $\mathbf{e}$, $\mathbf{Q}$, $\mathbf{q}$, ρ)

输入：$\mathcal{N}$——一个贝叶斯网；m——样本量；
　　$\mathbf{E}$——证据变量；$\mathbf{e}$——证据变量的取值；
　　$\mathbf{Q}$——查询变量；$\mathbf{q}$——查询变量的取值；
　　ρ——非证据变量的抽样顺序.
输出：对 $P(\mathbf{Q}=\mathbf{q}\mid\mathbf{E}=\mathbf{e})$ 的近似；

```
m_q ← 0 ;
随机生成一个与 E = e 一致的样本 D_1 ;
if (D_1 与 Q = q 一致)
    m_q ← m_q + 1 ;
end if
for (i = 2 to m)
    D_i ← D_{i-1} ;
    for (ρ 中的每一个变量 Z)
        设 Y = mb(Z), y_i 是 Y 在 D_i 中的当前取值，从 P(Z | Y = y_i) 抽样；
        用抽样结果替代 D_i 中 Z 的取值；
    end for
    if (D_i 与 Q=q 一致)
        m_q ← m_q + 1;
    end if
end for
return m_q/m.
```

图 6.3　吉布斯抽样算法

例 6.4　对如图 6.1 所示的贝叶斯网，用吉布斯抽样算法计算 $P(R=t\mid S=t)$. 首先随机生成一个与证据 $\{S=t\}$ 一致的样本，假设它是 $\mathbf{D}_1=\{C=t, R=t, S=t, W=f\}$. 接下来生成样本 $\mathbf{D}_2$：算法从 $\mathbf{D}_2=\mathbf{D}_1=\{C=t, R=t, S=t, W=f\}$ 出发，对非证据变量逐个抽样. 设抽样顺序为 $\langle C, R, W\rangle$. 抽样过程

如下：

(1) 对 C 进行抽样，抽样分布为 $P(C \mid R=t, S=t) \approx (0.444, 0.556)$，假设抽样结果为 $C=f$，于是 $\mathbf{D}_2$ 变为 $\{C=f, R=t, S=t, W=f\}$；

(2) 对 R 进行抽样，此时 C 的取值是 f，因此抽样分布为 $P(R \mid C=f, S=t, W=f) \approx (0.024, 0.976)$，假设抽样结果为 $R=f$，于是 $\mathbf{D}_2$ 变为 $\{C=f, R=f, S=t, W=f\}$；

(3) 对 W 进行抽样，此时 R 的取值是 f，因此抽样分布为 $P(W \mid R=f, S=t)=(0.9, 0.1)$，假设抽样结果为 $W=f$，于是 $\mathbf{D}_2$ 仍是 $\{C=f, R=t, S=t, W=f\}$，这是 $\mathbf{D}_2$ 的最终值. □

设抽样共得到 m 个样本，其中满足 $\mathbf{Q}=\mathbf{q}$ 的有 $m_{\mathbf{q}}$ 个. 那么，可以按下式近似计算后验概率：

$$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e}) \approx \frac{m_{\mathbf{q}}}{m}.$$

吉布斯抽样实际上是在贝叶斯网所有变量的联合状态空间中与 $\mathbf{E}=\mathbf{e}$ 一致的那个子空间里进行**随机漫步**（random walk）. 它先任意选择一个起点，以后的每一步都只依赖于前一步的状态，即上一个样本. 在数学上，这是一个**马尔可夫链**（Markov chain），简称**马氏链**. 它是一个离散时间随机过程，其基本性质可以通俗地理解为“给定现在，将来与过去无关”.

在满足一定条件的前提下，无论开始于哪一个状态，马氏链在第 i 步状态的分布在 $i \to \infty$ 时都将收敛于一个**平稳分布**（stationary distribution）. 吉布斯抽样算法中的马氏链所对应的平稳分布正好是 $P(\mathbf{Z} \mid \mathbf{E}=\mathbf{e})$，这里 $\mathbf{Z}$ 是所有非证据变量的集合. 因此当总样本数 m 趋于无穷时，吉布斯算法相当于从 $P(\mathbf{Z} \mid \mathbf{E}=\mathbf{e})$ 中抽样. 这就保证了它所得样本中 $\mathbf{Q}=\mathbf{q}$ 出现的频率收敛于 $P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e})$.

吉布斯抽样的缺点是收敛速度慢，因为马氏链往往需要花很长时间才能趋于平稳分布. 另外，当网络中存在极端概率 0 和 1 时，无法保证马氏链存在平稳分布. 此时，吉布斯抽样将给出错误的结果.

6.2 变分法

本节介绍一类直接针对解的精度进行优化的近似推理算法，即**变分法**（variational methods）①. 变分法的基本思想是通过**变分转换**（variational transformation），将概率推理问题转换为一个**变分优化**（variational optimization）问题

① 经典变分法主要讨论泛函极值问题，现代变分法已经成为一个重要的优化计算工具，被广泛应用于有限元分析、量子力学、统计力学以及优化控制等领域.

来处理. 对于比较困难的概率推理问题，对应的变分优化问题往往也缺乏有效的精确解法，此时可以对变分优化问题进行适当的放松，如对其目标函数和约束条件集进行简化，并求助于迭代的方法，以期能高效地获得一个近似解. 下面介绍变分法的两个例子，即朴素平均场法和循环传播法.

6.2.1 朴素平均场法

朴素平均场理论（naïve mean field theory）是统计力学常用的近似方法. 它关心的是在一个包含众多粒子的系统中，其它粒子对某个特定粒子的作用. 它的基本思想是用一个“平均了的场”来代替原来的系统，并在其中对上述作用进行估计. 这样，它把一个高自由度的复杂问题转化为一个低自由度的简单问题来近似处理. 把朴素平均场的基本思想应用于贝叶斯网推理，可以在拥有众多变量且连接稠密的网络中，近似计算后验概率.

设 $\mathcal{N}$ 为一贝叶斯网，表示联合概率分布 $P(\mathbf{X})$. 设观测到的证据为 $\mathbf{E}=\mathbf{e}$，所有非证据变量的集合为 $\mathbf{Z}=\{Z_1,Z_2,\cdots,Z_n\}$，准备计算的是 $P(\mathbf{E}=\mathbf{e})$ 和 $P(\mathbf{Z}\mid\mathbf{E}=\mathbf{e})$.

变分法首先把这个概率推理问题转化为变分优化问题. 定义如下目标函数：

$$J(Q)=\log P(\mathbf{E}=\mathbf{e})-KL[Q(\mathbf{Z}),P(\mathbf{Z}\mid\mathbf{E}=\mathbf{e})]. \tag{6.13}$$

这里 $Q(\mathbf{Z})$ 是一族用来近似 $P(\mathbf{Z}\mid\mathbf{E}=\mathbf{e})$ 的参数化分布，称为 **变分分布**（variational distribution），而 $KL[Q(\mathbf{Z}),P(\mathbf{Z}\mid\mathbf{E}=\mathbf{e})]$ 是变分分布 $Q(\mathbf{Z})$ 和待求的后验概率分布 $P(\mathbf{Z}\mid\mathbf{E}=\mathbf{e})$ 之间的 KL 距离. 注意，式（6.13）可以展开为

$$\begin{aligned}J(Q)&=\log P(\mathbf{E}=\mathbf{e})-\sum_{\mathbf{Z}}Q(\mathbf{Z})\log\frac{Q(\mathbf{Z})}{P(\mathbf{Z}\mid\mathbf{E}=\mathbf{e})}\\&=\log P(\mathbf{E}=\mathbf{e})-\sum_{\mathbf{Z}}Q(\mathbf{Z})\log\frac{Q(\mathbf{Z})P(\mathbf{E}=\mathbf{e})}{P(\mathbf{Z},\mathbf{E}=\mathbf{e})}\\&=-\sum_{\mathbf{Z}}Q(\mathbf{Z})\log Q(\mathbf{Z})+\sum_{\mathbf{Z}}Q(\mathbf{Z})\log P(\mathbf{Z},\mathbf{E}=\mathbf{e})\\&=H_Q(\mathbf{Z})+\sum_{\mathbf{Z}}Q(\mathbf{Z})\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}).\end{aligned} \tag{6.14}$$

式（6.14）中的第一项 $H_Q(\mathbf{Z})$ 是 $\mathbf{Z}$ 相对于变分分布 $Q(\mathbf{Z})$ 的熵.

考虑目标函数 $J(Q)$ 的最大值点 Q^* 的性质. 因为 KL 距离非负，且只有当它涉及到的两个分布相同时才取最小值 0，所以使 $J(Q)$ 取最大值的最优变分分布就是后验概率分布，即

$$Q^*(\mathbf{Z})=P(\mathbf{Z}\mid\mathbf{E}=\mathbf{e}).$$

此时，

$$J(Q^*)=\log P(\mathbf{E}=\mathbf{e}).$$

因此，只要解决了这个变分优化问题，就得到了原概率推理问题的精确解.

在实际中，上述变分优化问题的精确解并不容易得到. 一个根本原因就是变分分布 $Q(\mathbf{Z})$ 的空间难以显式表达. 因此，为了简化计算，朴素平均场法根据平均场理论的思想，将变分分布限制在一类最简单的分布上，即用所有变量都相互独立的分布来近似后验分布，即

$$Q(\mathbf{Z}) = \prod_{i=1}^{n} Q_i(Z_i). \tag{6.15}$$

尽管这是个很简单的分布，仍然有 $\sum_{i=1}^{n}(|Z_i|-1)$ 个独立参数可以调节. 将式 (6.15) 代入式 (6.14) 可得

$$J(Q) = \sum_{i=1}^{n} H_{Q_i}(Z_i) + \sum_{\mathbf{Z}}\Big[\prod_{i=1}^{n} Q_i(Z_i)\Big]\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}). \tag{6.16}$$

式 (6.16) 的最大值和最大值点一般没有闭公式解. 根据式 (6.15)，变分分布 $Q(\mathbf{Z})$ 的边缘分布可以独立地改变. 因此，可以利用如下迭代法来优化目标函数 $J(Q)$：

(1) 可以从某个初始分布 $Q^0(\mathbf{Z})$ 开始；

(2) repeat 直到收敛

　　for $i = 1$ 到 n

　　以 $J(Q)$ 为目标函数，对变分边缘分布 $Q_i(Z_i)$ 进行优化.

在优化 $Q_i(Z_i)$ 时，其它各变分边缘分布 $\{Q_j(Z_j) \mid j \neq i\}$ 均当做常数来处理，因此 $J(Q)$ 仅为 $Q_i(Z_i)$ 的函数，即

$$J(Q_i) = \mathscr{C} + H(Q_i) + \sum_{Z_i} Q_i(Z_i)\, E_Q[\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}) \mid Z_i], \tag{6.17}$$

其中，$\mathscr{C}$ 是与 $Q_i(Z_i)$ 无关的常数，而

$$\begin{aligned} E_Q[\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}) \mid Z_i] &= \sum_{\mathbf{Z}\setminus\{Z_i\}}\Big[\prod_{j=1,j\neq i}^{n} Q_j(Z_j)\Big]\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}) \\ &= \sum_{\mathbf{Z}\setminus\{Z_i\}} Q(\mathbf{Z}\setminus\{Z_i\} \mid Z_i)\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}) \end{aligned}$$

是函数 $\log P(\mathbf{Z},\mathbf{E}=\mathbf{e})$ 在分布 $Q(\mathbf{Z}\setminus\{Z_i\} \mid Z_i)$ 下的期望值.

注意，$\sum_{Z_i} Q(Z_i) = 1$. 为优化 $J(Q_i)$，引入**拉格朗日乘数** (Lagrange multiplier) λ，并求解如下方程组：

$$\frac{\partial}{\partial Q_i(Z_i = z)}\Big[J(Q_i) + \lambda\Big(\sum_{Z_i} Q(Z_i) - 1\Big)\Big] = 0,\ \forall z \in \Omega_{Z_i},$$

即

$$-\log Q_i(Z_i = z) + 1 + E_Q[\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}) \mid Z_i] + \lambda = 0.$$

由上式可得如下用于更新 $Q_i(Z_i)$ 的**平均场方程** (mean field equation)：

$$Q_i(Z_i) = \frac{1}{\alpha_i} e^{E_Q[\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}) \mid Z_i]}, \tag{6.18}$$

这里 α_i 是规一化常数，

$$\alpha_i = \sum_{Z_i} e^{E_Q[\log P(\mathbf{Z},\mathbf{E}=\mathbf{e}) \mid Z_i]}. \tag{6.19}$$

上述迭代过程的每一步都将使 $J(Q)$ 增大，因此它最终会收敛. 但是不一定会收敛到全局最优，初始点的选取和迭代更新的顺序都会影响到解的质量.

朴素平均场法直接对所求的后验概率分布的精度进行优化，概念清晰，算法的收敛速度比随机抽样法快. 当网络中变量数目多、连接稠密，但变量之间的依赖关系却比较弱的时候，它可以给出一个较好的结果. 但是，实际中变分平均场分布 $\prod_{i=1}^{n} Q_i(Z_i)$ 往往过于简单，不能很好地近似真实的后验分布.

6.2.2 循环传播算法

在贝叶斯网发展初期，Pearl（1986）提出了一个仅适用于底图而不含（无向）圈的网络的精确推理算法. 后来人们发现，如果把该算法在底图含圈的网络上强制运行，往往也可以得到一个较好的近似解，这就是**循环传播**（loopy propagation）算法，又称为**迭代信度传播**（iterative belief propagation）算法. 该算法的变种曾分别独立地出现于统计物理、计算机视觉以及信息编码等研究领域. 最近的研究还表明，它可以看成是变分法的一个特例. 下面先给出循环传播算法的一般流程，然后简单介绍它与变分法的关系.

循环传播算法的基本步骤是，先在网络中的各节点之间迭代地进行信息传播，然后利用网络中的信息近似计算任一节点的信度，即后验概率分布. 考虑贝叶斯网 $\mathcal{N}$ 中的任一节点 X，设其父节点为 $\mathbf{U} = \{U_1, U_2, \cdots, U_m\}$，子节点为 $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_n\}$. 在第 t 次迭代中，节点 X 向它的每个父节点 U_i 发送一个信息，这个信息是 U_i 的函数，记为 $\lambda_X^t(U_i)$；它同时还向它的每个子节点 Y_j 发送一个信息，这个信息是 X 的函数，记为 $\pi_{Y_j}^t(X)$，如图 6.4 所示. 当 $t = 0$ 时，$\lambda_X^0(U_i)$ 和 $\pi_{Y_j}^0(X)$ 是单位函数 1. 当 $t > 0$，它们由如下两式递归地给出：

$$\lambda_X^t(U_i) = \alpha \sum_X \lambda_X(X) \prod_j \lambda_{Y_j}^{t-1}(X) \sum_{\mathbf{U}\setminus\{U_i\}} P(X \mid \mathbf{U}) \prod_{k\neq i} \pi_X^{t-1}(U_k), \tag{6.20}$$

$$\pi_{Y_j}^t(X) = \alpha \lambda_X(X) \prod_{k\neq j} \lambda_{Y_k}^{t-1}(X) \sum_{\mathbf{U}} \mathrm{P}(\mathrm{X} \mid \mathbf{U}) \prod_i \pi_X^{t-1}(U_i), \tag{6.21}$$

其中，α 是一个规一化常数，而 $\lambda_X(X)$ 是 X 向自己发送的信息，它的作用是把观测证据 $\mathbf{E} = \mathbf{e}$ 也考虑进来：如果 $X \in \mathbf{E}$ 且观测到它的取值为 x_i，则

$$\lambda_X(X = x_j) = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases};$$

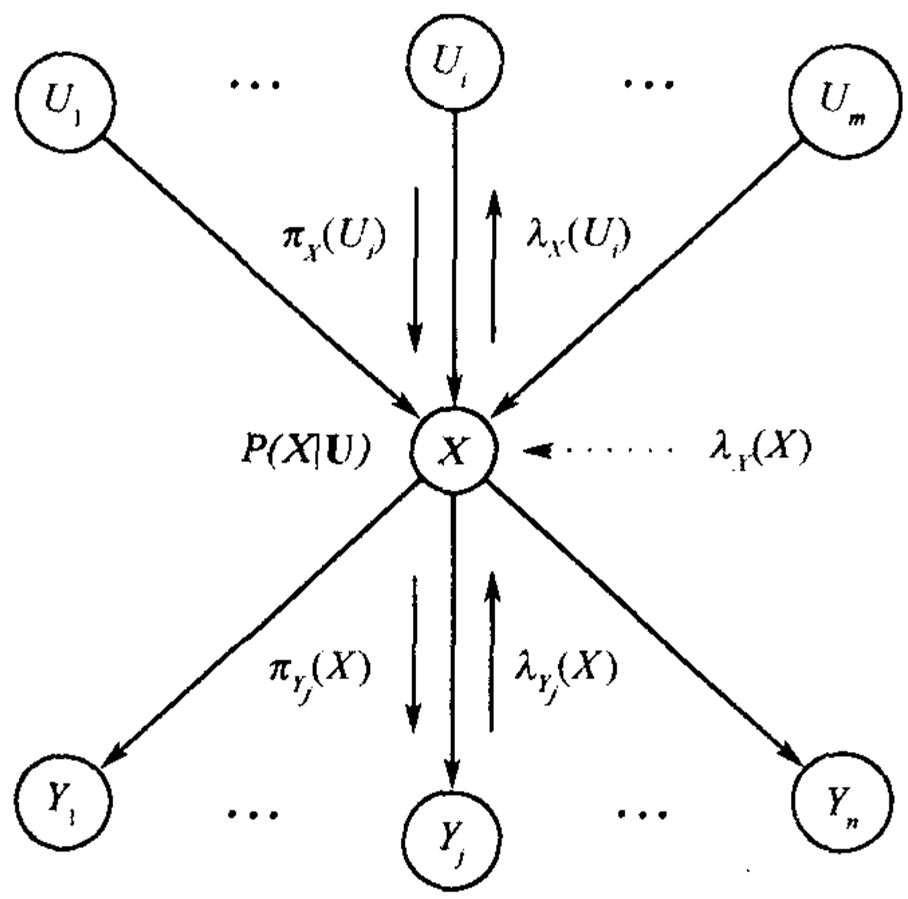


图 6.4 节点 X 处的信息传播示意

否则，$\lambda_X(X)$ 是单位函数 **1**.

在第 t 次迭代结束时，可以用网络中的信息计算变量 X 当前的信度：

$$\mathrm{Bel}^t(X) = \alpha\lambda^t(X)\pi^t(X). \tag{6.22}$$

这里，

$$\lambda^t(X) = \lambda_X(X)\prod_j \lambda^t_{Y_j}(X), \tag{6.23}$$

$$\pi^t(X) = \sum_{\mathbf{U}} P(X \mid \mathbf{U}) \prod_i \pi^t_X(U_i). \tag{6.24}$$

其中 $\lambda^t(X)$ 集中了所有来自 X 的子节点的影响，而 $\pi^t(X)$ 则 集中了所有来自 X 的父节点的影响.

当迭代次数 t 达到某个预先设定的上限，或者两次相邻迭代之间信度的变化不超过某个预先设定的阈值时，算法终止. 我们用此时 X 的信度来近似它的后验概率分布，即

$$P(X \mid \mathbf{E} = \mathbf{e}) \approx \mathrm{Bel}^t(X). \tag{6.25}$$

循环传播算法的性能与网络的结构密切相关. 对于底图不含圈的网络，循环传播算法能够迅速地收敛于真实的后验概率分布. 而对于底图含圈的网络，循环传播算法不能保证一定收敛，信息可能在网络中的回路内无限循环传播，形成振荡.

为什么说循环传播算法是变分法的一个特例呢？Yedidia 等（2000）指出，循环传播算法实际上是在对 Bethe 自由能（Bethe free energy）进行最小化. Wainwright 和 Jordan（2003）进一步指出，变分平均场法和循环传播法是对同一凸优化问题的两种不同的近似解法：前者将优化对象限制在一类简单分布上，

后者则相当于放松优化问题的约束.

6.3 其它近似推理算法

除随机抽样法和变分法之外，还有两类常用的近似推理算法，即**模型简化法**（model simplification methods）和**基于搜索的算法**（search-based algorithms）.

模型简化法的基本思想是先对模型进行简化，直到精确推理可以进行时，在简化后的模型上进行精确推理，得到近似解. 模型简化的方法有许多种，如删除依赖关系较弱的边（Kjaerulff，1994；van Engelen，1997）、除去影响较小的变量（Draper，1995）、缩小变量的状态空间（Wellman and Liu，1994）、将小概率值设为 0（Jensen and Andersen，1990）、用较为简单的模型来近似取代较为复杂的模型（Sarkar，1993）以及部分评估（Horvitz，1990）等. 变分平均场法也可以看做是一种模型简化方法，它用相对简单的变分分布来近似较为复杂的分布.

基于搜索的算法通常假设贝叶斯网的联合概率分布主要集中在变量联合状态空间的一个很小的子空间内. 算法寻找在这个子空间的状态组合，计算其概率，并用它们来近似联合概率分布，进行推理. 这类算法包括“Top-N”算法（Henrion，1988）、使用“冲突”（conflicts）进行搜索的算法（Poole，1993）、决定型近似（deterministic approximation）算法（Santos and Shimony，1998）、“抽样-累积”（sample-and-accumulate）算法（Santos et al.，1996），以及基于蚁群的算法（ant colony algorithms）的 MPE 近似算法（Guo et al.，2004）等. Druzdzel（1994）为基于搜索的算法提供了理论基础. 随机抽样算法也可以看做是基于搜索的算法，它用随机抽样的方法来搜索大概率事件，然后从所得样本出发进行近似推理.

贝叶斯网推理算法多种多样，一个很自然的问题是：对于一个给定的贝叶斯网，哪个推理算法性能最优？这就是所谓的**算法选择**问题. Jitnah 和 Nicholson（1996）研究了网络的特征和精确推理算法的性能之间的关系. Santos 等（1995）提出了一个综合多个推理算法长处的分布式概率推理体系结构. Guo（2003）提出了一个基于机器学习的 MPE 算法选择系统，能够实时选择适合给定网络的 MPE 近似算法.

6.4 文献介绍

重要性抽样法（Rubinstein，1981）是蒙特卡洛积分中降低方差的一种手段. Henrion（1988）提出了逻辑抽样，它是最简单也是最先被用于贝叶斯网近

似推理的重要性抽样算法. Fung 和 Chang（1989）、Shachter 和 Peot（1989）同时提出了似然加权算法. Shachter 和 Peot（1989）还提出了自重要性抽样和启发式重要性抽样算法. Fung 和 Favero（1994）提出了逆序抽样（backward sampling），它也是重要性抽样的一个特例. Cheng 和 Druzdzel（2000）提出了自适应重要性抽样算法，同时也给出了重要性抽样算法的通用框架.

MCMC 算法起源于统计物理中的近似计算（Metropolis et al.，1953）. 被广泛用于解决组合优化问题的模拟退火（simulated annealing）算法（Kirkpatrick et al.，1983）也是一种 MCMC 算法. Pearl（1987，1988）最先将吉布斯抽样算法应用于贝叶斯网推理. Neal（1993）给出了对 MCMC 算法的一个综述.

统计物理中的平均场理论最早曾被 Peterson 和 Anderson（1987）用于解决概率神经网络的推理和学习问题. 而变分法在贝叶斯网推理中的应用则主要是由 Jordan 的研究小组发展起来的（Saul et al.，1996；Jaakkola and Jordan，1999；Jordan et al.，1999；Saul and Jordan，2000；Wainwright and Jordan，2003）：Saul 等（1996）、Jaakkola 和 Jordan（1996）首先将其用于解决一类特殊的贝叶斯网，即 Sigmoid 网络的推理问题；Jaakkola 和 Jordan（1999）又将其应用于噪音或网络之中，解决了 QMR-DT 的近似推理问题. 循环传播算法在底图含圈的网络中并不总收敛于真实的后验概率分布（Murphy et al.，1999），但它在纠错码（Frey and Mackay，1998）和计算机视觉（Freeman et al.，2000）的研究中表现出色. Yedidia 等（Yedidia et al.，2001；Yedidia，2002；Yedidia et al.，2002）揭示了循环传播算法与自由能最小化问题的联系，说明了循环传播算法是变分法的一个特例. Wainwright 和 Jordan（2003）进一步提供了变分法的一个通用框架，在其中，变分平均场法和循环传播算法可以看做是对同一变分优化问题不同的近似解法.

第三部分

贝叶斯网学习

第7章 参数学习

贝叶斯网最初是作为一种处理专家系统中不确定性的工具而被提出的. 近年来，它越来越多地被用于数据分析，以揭示和刻画数据中所蕴含的规律. 贝叶斯网学习指的是通过分析数据而获得贝叶斯网的过程，它包括参数学习和结构学习两种情况. 参数学习指的是已知网络结构，确定网络参数的问题；结构学习则是既要确定网络结构，又要确定网络参数. 本章讨论参数学习，第 8、9 两章讨论结构学习.

7.1 贝叶斯网与数据分析

本节解释什么是数据，什么叫做用贝叶斯网来分析数据，以及一些相关概念.

设 $\mathbf{X}=\{X_1,X_2,\cdots,X_n\}$ 为一组随机变量. 由 $\mathbf{X}$ 中所有或部分变量的状态所构成的向量称为一个**数据样本** (data sample)①，简称为样本. 一些数据样本放在一起组成**数据组** (data collection)，它有时被简称为**数据**. 数据一般以表的形式给出. 表 7.1 给出了一组关于 5 个变量 $\{X_1,X_2,\cdots,X_5\}$ 的数据，其中每一列对应一个变量，每一行代表一个数据样本. 表中共有 4 个样本. 在前两个样本中，所有变量的状态都是已知的，所以称它们为**完整样本** (complete sample). 在后两个样本中，有些变量的状态未知，所以称它们为**缺值样本** (incomplete sample). 只包含完整样本的数据组称为**完整数据组** (complete data). 含有缺值样本的数据组则称为**缺值数据组** (incomplete data). 本书用黑体大写字母 $\mathbf{D}$ 表示一个数据样本，用花体字母 $\mathcal{D}$ 表示一组数据.

表 7.1 一组数据，其中“—”表示变量状态未知

X_1	X_2	X_3	X_4	X_5
1	1	2	2	1
1	1	2	1	2
1	—	2	1	1
—	1	2	2	—

① 有时也称为一个**数据例** (data case) 或**记录** (record).

用贝叶斯网来分析一组数据 $\mathscr{D}$，就是要从这组数据出发，找出一个相对于数据在某种意义下最优的贝叶斯网. 所得的结果是关于数据 $\mathscr{D}$ 的一个统计模型，称为**贝叶斯网模型**（Bayesian network model）.

正如第 2 章所述，一个贝叶斯网 $\mathscr{N}$ 有定性和定量两个方面的内容：定性内容包括变量之间的网络结构，记为 $\mathscr{G}$；定量内容则包括各变量的概率分布，记为 $\boldsymbol{\theta}$. 所以，$\mathscr{N}$ 又可以写成二元组（$\mathscr{G}$，$\boldsymbol{\theta}$）的形式. 在讨论数据分析时，$\mathscr{G}$ 被称为**模型结构**（model structure），有时简称**模型**①；而 $\boldsymbol{\theta}$ 则称为**模型参数**（model parameters）.

通过数据分析获得贝叶斯网模型的过程称为**贝叶斯网学习**（Bayesian network learning）. 当模型结构已知时，贝叶斯网学习又简称为**参数学习**（parameter learning）；而当模型结构未知时，它称为**结构学习**（structure learning）.

7.2　单参数最大似然估计

参数学习在统计学中称为**参数估计**（parameter estimation），有两种基本方法，即最大似然估计 和贝叶斯估计. 本节在单参数、完整数据的情况下介绍最大似然估计及其基本原理，7.3 节在同样的情况下介绍贝叶斯估计.

1. 最大似然估计的概念

我们从一个例子开始. 图 7.1（a）所示的是由一个随机变量 X 组成的简单贝叶斯网，设其中 X 表示投掷图钉的结果，它有两个可能的取值：h（头朝上）和 t（尾朝上），如图 7.1（b）所示. 这里，网络只有一个独立参数，即图钉头朝上的概率 $\theta = P(X=h)$②. 为估计 θ，我们投掷图钉多次. 设总共掷了 6 次，结果依次是 t、h、t、t、h、t. 基于这个结果，直观上可以用图钉头朝上出现的频率，即 $\frac{2}{6}$，作为对参数 θ 的估计.

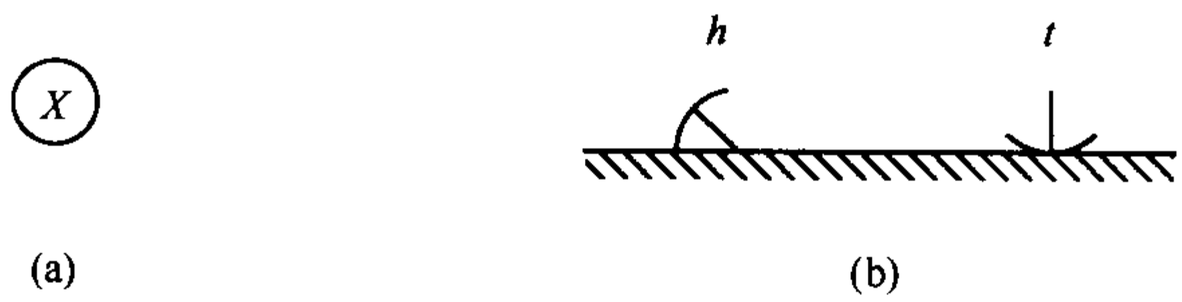


图 7.1　由一个随机变量 X 组成的简单贝叶斯网以及 X 的两个可能的取值

① 注意，“贝叶斯网模型”这个词既可指二元组（$\mathscr{G}$，$\boldsymbol{\theta}$），又可指其中的 $\mathscr{G}$. 这不会引起混淆，有必要时，我们会明示它指的是（$\mathscr{G}$，$\boldsymbol{\theta}$）还是 $\mathscr{G}$.

② 也可以选图钉尾朝上的概率作为独立参数.

上述直观思想的背后有没有什么基本原则？为什么可以用频率来估计概率参数？为回答此问题，考虑 θ 的另一个可能取值 0. 这个取值与观测到的数据 $\mathscr{D}=(t, h, t, t, h, t)$ 相矛盾：如果 $\theta=0$，那么“图钉头朝上”就应该不会出现，而实际上它出现了两次. 相比之下，$\theta=\frac{2}{6}$ 与数据则是相容的. 所以，如果要在 0 和 $\frac{2}{6}$ 两者之间选择一个作为对 θ 的估计，那么应该选 $\frac{2}{6}$. 这个例子说明，在判断某个数值作为 θ 的估计是否合适时，应该考虑它与数据的**拟合程度**（degree of fit）.

一般而言，参数 θ 的某个可能取值 θ_0 与数据 $\mathscr{D}$ 的拟合程度可以用数据的条件概率 $P(\mathscr{D}|\theta=\theta_0)$ 来度量. 这个概率越大，θ_0 与 $\mathscr{D}$ 的拟合程度就越高. 在上面的例子中，$P(\mathscr{D}|\theta=0)=0$，而 $P\left(\mathscr{D}|\theta=\frac{2}{6}\right)>0$，所以与 0 相比，$\frac{2}{6}$ 与 $\mathscr{D}$ 的拟合程度更高，因此更适合作为参数 θ 的估计.

给定 θ，数据 $\mathscr{D}$ 的条件概率 $P(\mathscr{D} \mid \theta)$ 称为是 θ 的**似然度**（likelihood），记为

$$L(\theta \mid \mathscr{D}) = P(\mathscr{D} \mid \theta). \tag{7.1}$$

如果固定 $\mathscr{D}$ 而让 θ 在其定义域上变动，那么 $L(\theta \mid \mathscr{D})$ 就是 θ 的一个函数，称为 θ 的**似然函数**（likelihood function）. 参数 θ 的**最大似然估计**（maximum likelihood estimation），简称 MLE，是令 $L(\theta \mid \mathscr{D})$ 达到最大的那个取值 θ^*，即

$$\theta^* = \arg\sup_{\theta} L(\theta \mid \mathscr{D}).^{①} \tag{7.2}$$

最大似然估计的概念明显可以推广到多参数和多变量的情况.

2. 最大似然估计的计算

设数据 $\mathscr{D}$ 由样本（$\mathbf{D}_1$，$\mathbf{D}_2$，…，$\mathbf{D}_m$）组成. 为了计算上的方便，需要做两个假设. 第一个假设是，$\mathscr{D}$ 中各样本在给定参数 θ 时相互独立，即

$$L(\theta \mid \mathscr{D}) = P(\mathscr{D} \mid \theta) = \prod_{i=1}^{m} P(\mathbf{D}_i \mid \theta). \tag{7.3}$$

第二个假设是，每个样本 $\mathbf{D}_i$ 的条件概率分布 $P(\mathbf{D}_i \mid \theta)$ 相同. 在掷图钉这个例子中，即设对每一个 i 都有

$$P(\mathbf{D}_i = h \mid \theta) = \theta,\ P(\mathbf{D}_i = t \mid \theta) = 1-\theta. \tag{7.4}$$

这两个假设是统计学中的基本假设，统称为**独立同分布**（independent and identically distributed）假设，简称 i. i. d. 假设.

---

① 这里我们假设 $L(\theta \mid \mathscr{D})$ 的最大值存在，且其最大值点唯一.

在投掷图钉的例子中，根据 i. i. d. 假设，有

$$L(\theta \mid \mathscr{D}) = P(\mathscr{D} \mid \theta) = \prod_{i=1}^{m} P(\mathbf{D}_i \mid \theta) = \theta^{m_h}(1-\theta)^{m_t}. \tag{7.5}$$

其中 m_h 和 m_t 分别是 $\mathscr{D}$ 中图钉头朝上和尾朝上出现的次数. 由于 m_h 和 m_t 概括了数据中所有关于似然函数的信息，所以称它们为**充分统计量**(sufficient statistics)①. 具有式(7.5)形式的似然函数称为**二项似然函数**(binomial likelihood function).

例 7.1 在本节开始提出的问题中，设 $\mathscr{D}=(t,h,t,t,h,t)$，有

$$\begin{aligned} L(\theta \mid \mathscr{D}) &= P(\mathscr{D} \mid \theta) \\ &= P(\mathbf{D}_1 = t \mid \theta)P(\mathbf{D}_2 = h \mid \theta)P(\mathbf{D}_3 = t \mid \theta)P(\mathbf{D}_4 = t \mid \theta) \\ &\quad P(\mathbf{D}_5 = h \mid \theta)P(\mathbf{D}_6 = t \mid \theta) \\ &= (1-\theta)\theta(1-\theta)(1-\theta)\theta(1-\theta) \\ &= \theta^2(1-\theta)^4. \end{aligned}$$

□

对似然函数 $L(\theta \mid \mathscr{D})$ 取对数，就得到**对数似然函数** (loglikelihood function)，即

$$l(\theta \mid \mathscr{D}) = \log L(\theta \mid \mathscr{D}). \tag{7.6}$$

显然，$l(\theta|\mathscr{D})$ 的最大值点也是 $L(\theta|\mathscr{D})$ 的最大值点，但前者在计算上更加方便. 在掷图钉的例子中，有

$$l(\theta \mid \mathscr{D}) = m_h \log\theta + m_t \log(1-\theta).$$

根据推论 1.1，$l(\theta|\mathscr{D})$ 的最大值点为

$$\theta^* = \frac{m_h}{m_h + m_t} = \frac{m_h}{m}. \tag{7.7}$$

其中 $m = m_h + m_t$ 是样本量. 这即是说，参数 θ 的最大似然估计就是数据中图钉头朝上出现的频率.

7.3 单参数贝叶斯估计

为引出贝叶斯估计的基本概念，考虑如下两个问题：

(1) 假设投掷一个图钉 6 次，其中 2 次头朝上，4 次尾朝上，问下一次投掷头朝上的概率多大?

(2) 假设投掷一枚硬币 6 次，其中 2 次正面向上，4 次反面向上，问下一次投掷正面向上的概率多大?

---

① 严格地说，数据 $\mathscr{D}$ 的函数 $S(\mathscr{D})$ 称为充分统计量，如果给定 $S(\mathscr{D})=s$，$\mathscr{D}$ 与参数 θ 独立，即 $P(\mathscr{D}|S(\mathscr{D})=s, \theta)= P(\mathscr{D}|S(\mathscr{D})=s)$.

按照最大似然估计原则，这两个问题的答案都应该是$\frac{2}{6}$. 但是，直观上它们却很不一样. 对于问题（1），由于我们很少有投掷图钉的经验，因此在投掷前没有关于其概率特性的先验知识，只能根据投掷结果来估计其概率. 所以，$\frac{2}{6}$作为问题（1）的答案是合理的. 对于问题（2），我们从日常生活中已经获得了关于投掷硬币的先验知识：如果投掷的是正常硬币而非特制的魔术硬币，则掷出正面和反面的概率应各为$\frac{1}{2}$. 在此问题中，由于只投掷了6次，其结果不足以使我们相信所用的是魔术硬币. 所以，问题（2）的答案应该是$\frac{1}{2}$，而不是$\frac{2}{6}$.

最大似然估计不能区分上述两种情况，因为它视待估参数为一个未知但固定的量，从而不能考虑先验知识的影响. 贝叶斯估计则视待估参数为一个随机变量，因而可以利用先验知识. 读者在后面会看到，贝叶斯估计对上面两个问题所给出的答案是不同的，符合直观，从而比最大似然估计更为合理.

1. 贝叶斯估计的概念

在贝叶斯估计的框架中，参数θ被视为随机变量，对它进行估计就是计算其后验概率分布. 为此，首先要选用一个概率分布$p(\theta)$来总结关于θ的先验知识，然后把数据$\mathscr{D}=(\mathbf{D}_1,\mathbf{D}_2,\cdots,\mathbf{D}_m)$的影响用似然函数$L(\theta|\mathscr{D})=P(\mathscr{D}|\theta)$来归纳，最后使用贝叶斯公式［式（1.17）］将先验分布和似然函数结合，得到θ的后验分布，即

$$p(\theta \mid \mathscr{D}) \propto p(\theta)L(\theta \mid \mathscr{D}). \tag{7.8}$$

这就是θ的**贝叶斯估计**（Bayesian estimation）.

最大似然估计的结果是一个固定的取值θ^*，根据它可以预测下一次投图钉得头朝上的概率是θ^*. 贝叶斯估计的结果是一个概率分布$p(\theta \mid \mathscr{D})$，它也可以用来进行预测. 实际上，下一次投掷得头朝上的概率为

$$\begin{aligned} P(\mathbf{D}_{m+1}=h \mid \mathscr{D}) &= \int P(\mathbf{D}_{m+1}=h,\theta \mid \mathscr{D})\mathrm{d}\theta \\ &= \int P(\mathbf{D}_{m+1}=h \mid \theta,\mathscr{D})p(\theta \mid \mathscr{D})\mathrm{d}\theta \\ &= \int \theta p(\theta \mid \mathscr{D})\mathrm{d}\theta. \end{aligned} \tag{7.9}$$

式（7.9）称为$P(\mathbf{D}_{m+1}=h|\mathscr{D})$的**完全贝叶斯估计**（full Bayesian estimation）.

计算完全贝叶斯估计所需的积分运算往往比较困难，所以有时也用$p(\theta|\mathscr{D})$的最大值点作为对$P(\mathbf{D}_{m+1}=h|\mathscr{D})$的估计，即

$$P(\mathbf{D}_{m+1}=h \mid \mathscr{D}) \approx \theta^{*}=\arg\sup_{\theta} p(\theta \mid \mathscr{D}). \tag{7.10}$$

这个估计称为 $P(\mathbf{D}_{m+1}=h \mid \mathscr{D})$ 的**贝叶斯 MAP 估计**（Bayesian MAP estimation）. 在下面提及贝叶斯估计时，若无特别说明，都指的是完全贝叶斯估计.

2. 贝叶斯估计的计算

与 7.2 节类似，为了计算方便，需要做一些假设. 首先要假设的是，给定 θ，$\mathscr{D}$ 中的样本满足 i. i. d. 假设. 由于这个假设，式（7.8）可以改写为

$$p(\theta \mid \mathscr{D}) \propto \theta^{m_h}(1-\theta)^{m_t} p(\theta). \tag{7.11}$$

其中 m_h 和 m_t 分别是数据 $\mathscr{D}$ 中图钉头朝上和尾朝上的样本个数. 关于先验分布 $p(\theta)$，一般假设它是个贝塔分布 $B[\alpha_h, \alpha_t]$，即

$$p(\theta)=\frac{\Gamma(\alpha_t+\alpha_h)}{\Gamma(\alpha_t)\Gamma(\alpha_h)}\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1} \tag{7.12}$$

其中 $\Gamma(\cdot)$ 是 Γ 函数①，α_h 和 α_t 是分布的两个参数，有时称为**超参数**（hyper parameter）. 图 7.2 给出了几个贝塔分布. 假设 $p(\theta)$ 为 $B[\alpha_h, \alpha_t]$ 实际上就是做如下假设：关于 θ 的先验知识相当于已投掷图钉 $\alpha_h+\alpha_t$ 次，其中 α_h 次头朝上，α_t 次尾朝上. 所以，$\alpha_h+\alpha_t$ 称为**等价样本量**（equivalent sample size）.

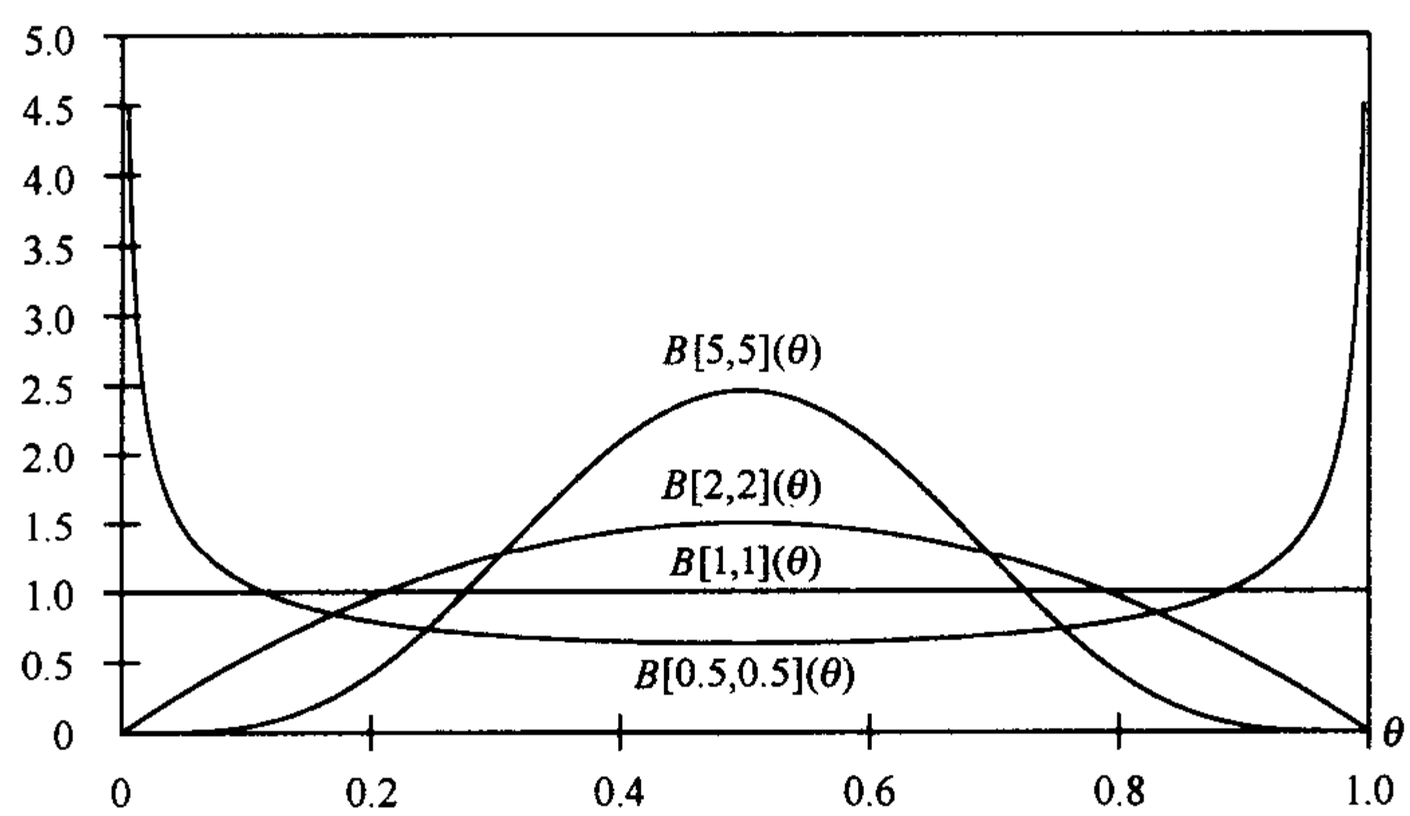


图 7.2 几个贝塔分布

将式（7.12）代入式（7.11），得

$$p(\theta \mid \mathscr{D}) \propto \theta^{m_h+\alpha_h-1}(1-\theta)^{m_t+\alpha_t-1}. \tag{7.13}$$

也就是说，θ 的后验分布 $p(\theta \mid \mathscr{D})$ 也是贝塔分布，其超参数为 $m_h+\alpha_h$ 和 $m_t+\alpha_t$，即

① $\Gamma(x)=\int_0^{\infty} t^{x-1}\mathrm{e}^{-t}\mathrm{d}t$. 若 x 为正整数，则 $\Gamma(x)$ 简化为 $(x-1)!$.

$$p(\theta \mid \mathscr{D}) = \frac{\Gamma(m_h + \alpha_h + m_t + \alpha_t)}{\Gamma(m_h + \alpha_h)\Gamma(m_t + \alpha_t)} \theta^{m_h + \alpha_h - 1}(1-\theta)^{m_t + \alpha_t - 1}. \tag{7.14}$$

下面考虑下一次投掷图钉得到头朝上的概率 $P(\mathbf{D}_{m+1} = h \mid \mathscr{D})$ 的完全贝叶斯估计. 根据式 (7.9) 和式 (7.14)，有

$$\begin{aligned} P(\mathbf{D}_{m+1} = h \mid \mathscr{D}) &= \int \theta p(\theta \mid \mathscr{D})\, \mathrm{d}\theta \\ &= \frac{\Gamma(m_h + \alpha_h + m_t + \alpha_t)}{\Gamma(m_h + \alpha_h)\ \Gamma(m_t + \alpha_t)} \int \theta\, \theta^{m_h + \alpha_h - 1}(1-\theta)^{m_t + \alpha_t - 1}\, \mathrm{d}\theta \\ &= \frac{m_h + \alpha_h}{m_h + m_t + \alpha_h + \alpha_t} \\ &= \frac{m_h + \alpha_h}{m + \alpha}. \end{aligned} \tag{7.15}$$

其中 $m = m_h + m_t$，$\alpha = \alpha_h + \alpha_t$. 注意，当样本量 m 很小的时候，这个估计主要依赖于先验知识；当样本量 m 增大时，这个估计越来越多地依赖于数据，越来越接近最大似然估计 $\frac{m_h}{m}$，而先验知识的影响逐渐减小.

例 7.2 考虑本节开始时提出的问题 (1). 由于没有关于投掷图钉的先验知识，所以假设先验分布是贝塔分布 $B[0, 0]$，于是，$\alpha_h = \alpha_t = 0$. 而 $m_h = 2$，$m_t = 4$. 根据式 (7.15)，下一次投掷图钉得到头朝上的概率为

$$\frac{2+0}{6+0} = \frac{2}{6}.$$

□

例 7.3 考虑本节开始时提出的问题 (2). 由于知道投掷正常硬币正面朝上的概率为 $\frac{1}{2}$，而且问题所涉及的硬币应该是正常硬币，但又不能完全排除它是魔术硬币的可能性，所以假设先验分布是贝塔分布 $B[100, 100]$，于是，$\alpha_h = \alpha_t = 100$. 而 $m_h = 2, m_t = 4$. 根据贝叶斯估计，下一次投得硬币正面朝上的概率为

$$\frac{2+100}{6+200} \approx \frac{1}{2}.$$

这里样本量 $m = 6$ 很小，先验知识占主导作用.

另一方面，如果假设共投掷硬币 6 万次，得 2 万次正面朝上，4 万次反面朝上，即 $m_h = 20000$，$m_t = 40000$. 那么，下一次投得硬币正面朝上的概率是

$$\frac{20000+100}{60000+200} \approx \frac{2}{6}.$$

这里样本量 $m = 60000$ 很大，数据占主导作用. □

3. 共轭分布族

为加深理解，我们再来看一看式 (7.8). 它的右边是先验概率分布 $p(\theta)$

和似然函数 $L(\theta \mid \mathscr{D})$ 的乘积. 在 i. i. d. 假设下，$L(\theta \mid \mathscr{D})$ 是二项似然函数. 上面假设先验分布 $p(\theta)$ 来自贝塔分布族. 这是因为贝塔分布族是二项似然函数的**共轭分布族**（conjugate family），即如果先验分布 $p(\theta)$ 是贝塔分布，那么后验分布 $p(\theta \mid \mathscr{D})$ 也是贝塔分布. 这使得贝叶斯估计的计算简单易行. 事实上，如果假设 $p(\theta)$ 来自另一分布族，比方说正态分布，那么贝叶斯估计计算起来就要困难得多.

另外，共轭分布族的使用也使得我们可以清楚地了解到贝叶斯估计是怎样把先验知识与观测数据结合到一起的. 假设 $p(\theta)$ 为贝塔分布 $B[\alpha_h, \alpha_t]$ 实际上就是做如下假设：先验知识相当于一组包含 α_h 个 头朝上和 α_t 个尾朝上的样本的**虚拟数据**（imaginary data）. 贝叶斯估计把这些虚拟数据和实际观测所得到的数据放到一起，得到一组由 $m_h + \alpha_h$ 个头朝上和 $m_t + \alpha_t$ 个尾朝上的样本所组成的数据. 于是，$p(\theta \mid \mathscr{D})$ 是 $B[m_h + \alpha_h, m_t + \alpha_t]$.

4. 顺序学习与批量学习

机器学习有两种模式，即顺序学习和批量学习. **顺序学习**（sequential learning）指一个一个地处理数据样本，每处理一个样本就更新一次参数，而且更新是在当前参数值的基础上进行的. **批量学习**（batch learning）则指同时处理所有数据，一次性得到参数估计. 在处理完当前数据之后的一段时间内，如果有新的数据出现，就把新老数据混合在一起，重新进行参数估计，这个过程完全不依赖于以前的估计. 贝叶斯估计既可以用于顺序学习，又可以用于批量学习，而最大似然估计只能用于批量学习.

7.4 单变量网络参数估计

前两节针对单参数贝叶斯网介绍了最大似然估计和贝叶斯估计，接下来的几节将把这两个概念和相关结果逐步推广.

本节考虑由一个多值变量 X 组成的贝叶斯网. 设 X 有 r 个取值，$\Omega_X = \{x_1, x_2, \cdots, x_r\}$. 网络的参数包括 $\theta_i = P(X = x_i), i = 1, 2, \cdots, r$. 用 $\boldsymbol{\theta}$ 记向量 $(\theta_1, \theta_2, \cdots, \theta_r)$. 由于概率分布的规范性，$\sum_{i=1}^{r} \theta_i = 1$，实际上该网络只有 $r-1$ 个独立参数.

先考虑 $\boldsymbol{\theta}$ 的最大似然估计. 设有一组 i. i. d. 数据 $\mathscr{D} = (\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m)$，其中满足 $X = x_i$ 的样本个数是 m_i，则有

$$L(\boldsymbol{\theta} \mid \mathscr{D}) = \prod_{i=1}^{r} \theta_i^{m_i}. \tag{7.16}$$

具有这个形式的似然函数称为**多项似然函数**（multinomial likelihood function），其中的 $\{m_i \mid i=1,2,\cdots,r\}$ 是充分统计量. 相应的对数似然函数为

$$l(\boldsymbol{\theta} \mid \mathscr{D}) = \sum_{i=1}^{r} m_i \log\theta_i. \tag{7.17}$$

根据推论 1.1，$\boldsymbol{\theta}$ 的最大似然估计 $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \cdots, \theta_r^*)$ 由下式给出：

$$\theta_i^* = \frac{m_i}{m}, \tag{7.18}$$

其中 $m = \sum_{i=1}^{r} m_i$ 是样本量.

接下来考虑 $\boldsymbol{\theta}$ 的贝叶斯估计. 多项似然函数的共轭分布族是狄利克雷(Dirichlet)分布族. 所以，一般假设先验分布 $p(\boldsymbol{\theta})$ 是狄利克雷分布 $D[\alpha_1, \alpha_2, \cdots, \alpha_r]$，即

$$p(\boldsymbol{\theta}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{r}\Gamma(\alpha_i)} \prod_{i=1}^{r} \theta_i^{\alpha_i - 1}, \tag{7.19}$$

其中 $\alpha = \sum_{i=1}^{r} \alpha_i$. 注意当 $r=2$ 时，狄利克雷分布 $D[\alpha_1, \alpha_2]$ 就是 β 分布 $B[\alpha_1, \alpha_2]$. 假设 $p(\boldsymbol{\theta})$ 为狄利克雷分布 $D[\alpha_1, \alpha_2, \cdots, \alpha_r]$ 就等于假设关于 $\boldsymbol{\theta}$ 的先验知识相当于 α 个虚拟数据样本，其中满足 $X = x_i$ 的样本数为 α_i. 所以，α 称为等价样本量.

由于式（7.16）和式（7.19），$\boldsymbol{\theta}$ 的后验分布 $p(\boldsymbol{\theta} \mid \mathscr{D})$ 为

$$p(\boldsymbol{\theta} \mid \mathscr{D}) \propto \prod_{i=1}^{r} \theta_i^{m_i + \alpha_i - 1}. \tag{7.20}$$

这即是说，$p(\boldsymbol{\theta} \mid \mathscr{D})$ 是狄利克雷分布 $D[m_1+\alpha_1, m_2+\alpha_2, \cdots, m_r+\alpha_r]$.

关于下一个样本 $\mathbf{D}_{m+1}$ 的概率分布的贝叶斯估计如下：

$$P(\mathbf{D}_{m+1} = x_i \mid \mathscr{D}) = \int \theta_i p(\boldsymbol{\theta} \mid \mathscr{D}) \mathrm{d}\boldsymbol{\theta} = \frac{m_i + \alpha_i}{m + \alpha}. \tag{7.21}$$

与 7.3 节的情况类似，当样本量 m 很小时，这个估计主要依赖于先验知识；当样本量 m 增大时，这个估计越来越多地依赖于数据，越来越接近最大似然估计 $\frac{m_i}{m}$，而先验知识的影响越来越小.

7.5 一般网络最大似然估计

本节及下面的 3 节讨论由多个变量组成的一般贝叶斯网的参数估计. 本节讨论完整数据情况下的最大似然估计，7.6 节讨论完整数据情况下的贝叶斯估计，

第 7.7 节讨论缺值数据情况下的最大似然估计，最后第 7.8 节讨论缺值数据情况下的贝叶斯估计.

7.5.1 最大似然估计的计算

考虑一个由 n 个变量 $\mathbf{X}=\{X_1, X_2, \cdots, X_n\}$ 组成的贝叶斯网 $\mathcal{N}$. 不失一般性，设其中的节点 X_i 共有 r_i 个取值 1，2，…，r_i，其父节点 $\pi(X_i)$ 的取值共有 q_i 个组合，1，2，…，q_i. 若 X_i 无父节点，则 $q_i=1$. 那么，网络的参数为

$$\theta_{ijk}=P(X_i=k \mid \pi(X_i)=j), \tag{7.22}$$

其中 i 的取值范围是 $1\sim n$，而对一个固定的 i，j 和 k 的取值范围分别是从 $1\sim q_i$ 及 从 $1\sim r_i$. 用 $\boldsymbol{\theta}$ 记所有 θ_{ijk} 组成的向量. 注意，这些参数不是相互独立的，由于概率分布的规范性，有

$$\sum_{k=1}^{r_i}\theta_{ijk}=\sum_{k=1}^{r_i}P(X_i=k \mid \pi(X_i)=j)=1, \quad \forall i,j. \tag{7.23}$$

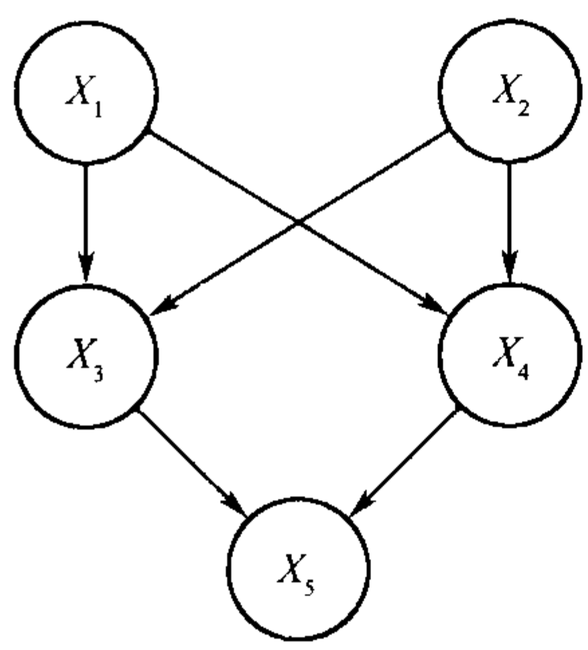


图 7.3 一个贝叶斯网

所以，$\mathcal{N}$ 的独立参数的个数是 $\sum_{i=1}^{n} q_i(r_i-1)$.

例 7.4 在图 7.3 所示的贝叶斯网中，设所有变量都取两个值，1 或 2. 将变量父节点的 取值组合按字典顺序排列，并从 1 开始编号. 例如，X_3 有两个父节点 X_1 和 X_2，它们的取值共有 4 种组合，排序为 $(X_1=1, X_2=1)$，$(X_1=1, X_2=2)$，$(X_1=2, X_2=1)$，$(X_1=2, X_2=2)$，编号分别是 1,2,3,4，则网络的参数如下：

$\theta_{111}=P(X_1=1)$, $\theta_{112}=P(X_1=2)$,

$\theta_{211}=P(X_2=1)$, $\theta_{212}=P(X_2=2)$,

$\theta_{311}=P(X_3=1 \mid X_1=1, X_2=1)$, $\theta_{312}=P(X_3=2 \mid X_1=1, X_2=1)$,

$\theta_{321}=P(X_3=1 \mid X_1=1, X_2=2)$, $\theta_{322}=P(X_3=2 \mid X_1=1, X_2=2)$,

$\theta_{331}=P(X_3=1 \mid X_1=2, X_2=1)$, $\theta_{332}=P(X_3=2 \mid X_1=2, X_2=1)$,

$\theta_{341}=P(X_3=1 \mid X_1=2, X_2=2)$, $\theta_{342}=P(X_3=2 \mid X_1=2, X_2=2)$,

…

□

设 $\mathcal{D}=(\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m)$ 是一组关于 $\mathcal{N}$ 的 i. i. d. 完整数据，则 $\boldsymbol{\theta}$ 的对数似然函数为

$$l(\boldsymbol{\theta} \mid \mathcal{D})=\log\prod_{l=1}^{m}P(\mathbf{D}_l \mid \boldsymbol{\theta})=\sum_{l=1}^{m}\log P(\mathbf{D}_l \mid \boldsymbol{\theta}).$$

为了得到关于 $\log P(\mathbf{D}_l \mid \boldsymbol{\theta})$ 的表达式，定义样本 $\mathbf{D}_l$ 的**特征函数**（characteristic function）$\chi(i, j, k:\mathbf{D}_l)$ 如下：

$$\chi(i, j, k:\mathbf{D}_l) = \begin{cases} 1, & \text{若在 } \mathbf{D}_l \text{ 中 } X_i = k \text{ 且 } \pi(X_i) = j \\ 0, & \text{若否} \end{cases}. \tag{7.24}$$

那么有

$$\log P(\mathbf{D}_l \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k:\mathbf{D}_l) \log \theta_{ijk}. \tag{7.25}$$

为了加深对这一关系的理解，下面来看一个例子.

例 7.5 考虑关于图 7.3 所示的贝叶斯网的样本 $\mathbf{D}_1=(1, 1, 2, 2, 1)$. 有

$\chi(1,1,1:\mathbf{D}_1)=\chi(2,1,1:\mathbf{D}_1)=\chi(3,1,2:\mathbf{D}_1)=\chi(4,1,2:\mathbf{D}_1)=\chi(5,4,1:\mathbf{D}_1)=1$,

而对其它的 i, j, k,则有

$$\chi(i,j,k:\mathbf{D}_1) = 0.$$

于是

$$\begin{aligned}
\log P(\mathbf{D}_1 \mid \boldsymbol{\theta}) &= \log P(X_1 = 1, X_2 = 1, X_3 = 2, X_4 = 2, X_5 = 1 \mid \boldsymbol{\theta}) \\
&= \log P(X_1 = 1 \mid \boldsymbol{\theta})\, P(X_2 = 1 \mid \boldsymbol{\theta})\, P(X_3 = 2 \mid X_1 = 1, X_2 = 1, \boldsymbol{\theta}) \\
&\quad P(X_4 = 2 \mid X_1 = 1, X_2 = 1, \boldsymbol{\theta})\, P(X_5 = 1 \mid X_3 = 2, X_4 = 2, \boldsymbol{\theta}) \\
&= \log \theta_{111}\, \theta_{211}\, \theta_{312}\, \theta_{412}\, \theta_{541} \\
&= \sum_{i=1}^{5} \sum_{j=1}^{q_i} \sum_{k=1}^{2} \chi(i,j,k:\mathbf{D}_1) \log \theta_{ijk}.
\end{aligned}$$

□

我们继续推导 $l(\boldsymbol{\theta} \mid \mathscr{D})$ 的表达式. 定义

$$m_{ijk} = \sum_{l=1}^{m} \chi(i, j, k:\mathbf{D}_l), \tag{7.26}$$

即 m_{ijk} 是数据中满足 $X_i = k$ 和 $\pi(X_i) = j$ 的样本的数量. 于是

$$\begin{aligned}
l(\boldsymbol{\theta} \mid \mathscr{D}) &= \sum_{l=1}^{m} \log P(\mathbf{D}_l \mid \boldsymbol{\theta}) \\
&= \sum_{l=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k:\mathbf{D}_l) \log \theta_{ijk} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \sum_{l=1}^{m} \chi(i, j, k:\mathbf{D}_l) \log \theta_{ijk} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \log \theta_{ijk}.
\end{aligned} \tag{7.27}$$

这里 $\{m_{ijk} \mid i=1,\cdots,n; j=1,\cdots,q_i; k=1,\cdots,r_i\}$ 是充分统计量. 对于任意固定的 i 和 j，由于 $\sum_{k=1}^{r_i}\theta_{ijk}=1$，根据推论 1.1，当 θ_{ijk} 取如下的值时①

$$\theta_{ijk}^{*}=\begin{cases}\dfrac{m_{ijk}}{\sum_{k=1}^{r_i} m_{ijk}}, & 若\sum_{k=1}^{r_i} m_{ijk}>0\\ \dfrac{1}{r_i}, & 若否,\end{cases} \tag{7.28}$$

表达式 $\sum_{k=1}^{r_i} m_{ijk}\log\theta_{ijk}$ 的值达到最大，从而 $l(\boldsymbol{\theta}\mid\mathscr{D})$ 达到最大. 所以，式 (7.28) 所给出的 θ_{ijk}^{*} 是 θ_{ijk} 的最大似然估计. 直观上有

$$\theta_{ijk}^{*}=\frac{\mathscr{D}\text{中满足 } X_i=k \text{ 和 } \pi(X_i)=j \text{ 的样本的数目}}{\mathscr{D}\text{中满足 } \pi(X_i)=j \text{ 的样本的数目}}.$$

例 7.6 考虑图 7.4(a) 所示贝叶斯网 $\mathscr{N}$，其中所有变量均取二值，1 或 2. 设图 7.4 (b) 是关于 $\mathscr{N}$ 的一组 i. i. d. 数据. 根据式 (7.28)，$\mathscr{N}$ 的参数的 MLE 估计分别如下：

$P(X_1)$

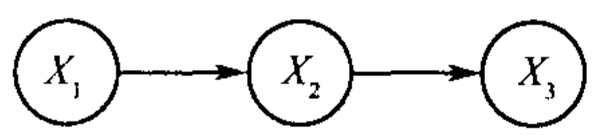
X_1	1	2
	2/4	2/4

$P(X_2 \mid X_1)$

X_1 \ X_2	1	2
1	2/2	0
2	0	2/2

$P(X_2 \mid X_1)$

X_2 \ X_3	1	2
1	1/2	1/2
2	0	2/2



(a) 贝叶斯网 $\mathscr{N}$

	X_1	X_2	X_3
$\mathbf{D}_1$	1	1	1
$\mathbf{D}_2$	2	2	2
$\mathbf{D}_3$	1	1	2
$\mathbf{D}_4$	2	2	2

(b) 完整数据

图 7.4 一个贝叶斯网 $\mathscr{N}$ 以及关于它的一组数据

□

① 当 $\sum_{k=1}^{r_i} m_{ijk}=0$ 时，与之相应的 $\theta_{ijk}(k=1,\cdots,r_i)$ 均不出现在 $l(\boldsymbol{\theta}\mid\mathscr{D})$ 中. 此时，无论 θ_{ijk} 取何值都不影响 $l(\boldsymbol{\theta}\mid\mathscr{D})$ 达到最大值. 因此可以设其为均匀分布.

7.5.2 最大似然估计的性质

下面讨论最大似然估计的性质. 设 $\mathcal{N}=(\mathcal{G},\boldsymbol{\theta}_{\mathcal{N}})$ 为一贝叶斯网. 用 $P_{\mathcal{N}}(\mathbf{X})$ 记 $\mathcal{N}$ 所表示的联合概率分布. 在 $\mathcal{N}$ 中，把参数换成最大似然估计 $\boldsymbol{\theta}^*$，得到另一个贝叶斯网 $\mathcal{N}^*=(\mathcal{G},\boldsymbol{\theta}^*)$，记其联合概率分布为 $P^*(\mathbf{X})$. 我们要回答如下两个问题：

(1) 如果数据 $\mathcal{D}$ 是从 $P_{\mathcal{N}}(\mathbf{X})$ 中抽样而得来的，那么当样本量 m 趋于无穷时，P^* 会不会收敛？会不会收敛到 $P_{\mathcal{N}}$？

(2) 如果数据 $\mathcal{D}$ 是从另一个分布 $P(\mathbf{X})$ $(P\neq P_{\mathcal{N}})$ 中抽样而得来的，那么当样本量 m 趋于无穷时，P^* 会不会收敛？收敛到什么分布？

如果数据 $\mathcal{D}$ 是从贝叶斯网 $\mathcal{N}$ 的联合分布 $P_{\mathcal{N}}(\mathbf{X})$ 中抽样而得来的，则称 $\mathcal{N}$ 是 $\mathcal{D}$ 的**原生模型**（generative model），称 $P_{\mathcal{N}}(\mathbf{X})$ 是 $\mathcal{D}$ 的**原生分布**（generative distribution）. 这个概念在讨论学习算法的性质时将经常用到.

引理 7.1 设 $\hat{P}(\mathbf{X})$ 是基于数据组 $\mathcal{D}$ 的**经验分布**（empirical distribution），即对 $\mathbf{X}$ 的任一取值 $\mathbf{x}$，

$$\hat{P}(\mathbf{X}=\mathbf{x})=\frac{\mathcal{D}\text{ 中满足 }\mathbf{X}=\mathbf{x}\text{ 的样本的数目}}{\text{样本量 }m}.$$

则当 $\hat{P}(\pi(X_i)=j)>0$ 时，

$$P^*(X_i=k\mid\pi(X_i)=j)=\theta^*_{ijk}=\hat{P}(X_i=k\mid\pi(X_i)=j).$$

证明 根据式(7.28)可知，在 $\hat{P}(\pi(X_i)=j)>0$ 时 θ^*_{ijk} 就是 $\hat{P}(X_i=k\mid\pi(X_i)=j)$. 引理得证. □

先考虑上述第二个问题. 对任一变量 X_i，它的一个取值 k，以及它在 $\mathcal{G}$ 中的父节点取值的一个组合 j，定义

$$\theta^{\mathcal{G}}_{ijk}=\begin{cases}P(X_i=k\mid\pi(X_i)=j), & \text{若 } P(\pi(X_i)=j)>0\\ \dfrac{1}{r_i}, & \text{若否}\end{cases}.$$

这个定义把概率分布 $P(\mathbf{X})$ 和模型结构 $\mathcal{G}$ 联系起来. 用 $P^{\mathcal{G}}(\mathbf{X})$ 记 $(\mathcal{G},\boldsymbol{\theta}^{\mathcal{G}})$ 的联合概率分布.

定理 7.1 当样本量 m 趋于无穷时，$P^*(\mathbf{X})$ 几乎必然收敛于 $P^{\mathcal{G}}(\mathbf{X})$，即

$$P^*(\mathbf{X})\xrightarrow[m\to\infty]{\text{a. s.}}P^{\mathcal{G}}(\mathbf{X}). \tag{7.29}$$

证明 由于 P^* 和 $P^{\mathcal{G}}$ 均是 $\mathcal{G}$-可分解的，只需证明对任意 i,j,k，有

$$P^*(X_i=k\mid\pi(X_i)=j)\xrightarrow[m\to\infty]{\text{a. s.}}P^{\mathcal{G}}(X_i=k\mid\pi(X_i)=j).$$

为了简洁，我们只考虑 $P(\pi(X_i)=j)>0$ 且 $\hat{P}(\pi(X_i)=j)>0$ 的情况. 使用引

理 7.1，得

$$P^*(X_i = k \mid \pi(X_i) = j) = \hat{P}(X_i = k \mid \pi(X_i) = j).$$

根据强大数定理，对 $\mathbf{X}$ 的任意取值 $\mathbf{x}$，有

$$\hat{P}(\mathbf{X} = \mathbf{x}) \xrightarrow[m\to\infty]{\text{a. s.}} P(\mathbf{X} = \mathbf{x}).$$

所以，

$$\begin{aligned}\hat{P}(X_i = k \mid \pi(X_i) = j) &\xrightarrow[m\to\infty]{\text{a. s.}} P(X_i = k \mid \pi(X_i) = j)\\ &= P^{\mathscr{G}}(X_i = k \mid \pi(X_i) = j).\end{aligned}$$

定理得证. □

定理 7.1 回答了前面提出的第二个问题，即当样本量趋于无穷时，$P^*(\mathbf{X})$ 收敛到 $P^{\mathscr{G}}(\mathbf{X})$. 那么，$P^{\mathscr{G}}(\mathbf{X})$ 与 $P(\mathbf{X})$ 有什么关系呢？下面的定理回答了这个问题.

定理 7.2 设 $Q^{\mathscr{G}}(\mathbf{X})$ 是一个 $\mathscr{G}$-可分解的概率分布，则有

$$\mathrm{KL}(P, P^{\mathscr{G}}) \leqslant \mathrm{KL}(P, Q^{\mathscr{G}}). \tag{7.30}$$

即在 Kullback-Leibler 距离的意义下，$P^{\mathscr{G}}$ 是与 P 最接近的 $\mathscr{G}$-可分解概率分布.

证明 由 KL 距离的定义［式（1.29）］，只需证明

$$\sum_{\mathbf{X}} P(\mathbf{X})\log P^{\mathscr{G}}(\mathbf{X}) \geqslant \sum_{\mathbf{X}} P(\mathbf{X})\log Q^{\mathscr{G}}(\mathbf{X}).$$

我们证明一个更强的结果，那就是上式对 $\mathscr{G}$ 中的任何一个祖先闭集 $\mathbf{Y}$ 成立，即

$$\sum_{\mathbf{Y}} (P(\mathbf{Y})\log P^{\mathscr{G}}(\mathbf{Y}) \geqslant \sum_{\mathbf{Y}} P(\mathbf{Y})\log Q^{\mathscr{G}}(\mathbf{Y}). \tag{7.31}$$

对 $\mathbf{Y}$ 中的变量个数 $|\mathbf{Y}|$ 做归纳. 当 $|\mathbf{Y}| = 1$ 时，$\mathbf{Y}$ 中只有一个节点，它在 $\mathscr{G}$ 中是根节点. 因此，$P(\mathbf{Y}) = P^{\mathscr{G}}(\mathbf{Y})$，从而根据信息不等式，式（7.31）必然成立.

假设式（7.31）在 $|\mathbf{Y}| = n$ 时成立. 考虑 $|\mathbf{Y}| = n+1$ 的情况. 设 Y 是 $\mathbf{Y}$ 中的一个不是 $\mathbf{Y}$ 中其它节点之父节点的节点，并设 $\mathbf{Y}' = \mathbf{Y} \setminus \{Y\}$. 为了简洁，我们只考虑 $P(\pi(Y)) > 0$ 的情况. 有

$$\begin{aligned}\sum_{\mathbf{Y}} P(\mathbf{Y})\log P^{\mathscr{G}}(\mathbf{Y}) &= \sum_{\mathbf{Y}} P(\mathbf{Y})\log P^{\mathscr{G}}(\mathbf{Y}') + \sum_{\mathbf{Y}} P(\mathbf{Y})\log P^{\mathscr{G}}(Y \mid \mathbf{Y}')\\ &= \sum_{\mathbf{Y}'} P(\mathbf{Y}')\log P^{\mathscr{G}}(\mathbf{Y}')\\ &\quad + \sum_{\pi(Y)} P[\pi(Y)] \sum_{Y} P(Y \mid \pi(Y))\log P^{\mathscr{G}}(Y \mid \pi(Y)).\end{aligned}$$

由归纳假设，有

$$\sum_{\mathbf{Y}'} P(\mathbf{Y}')\log P^{\mathscr{G}}(\mathbf{Y}') \geqslant \sum_{\mathbf{Y}'} P(\mathbf{Y}')\log Q^{\mathscr{G}}(\mathbf{Y}').$$

而 $P^{\mathscr{G}}(Y \mid \pi(Y)) = P(Y \mid \pi(Y))$. 由信息不等式，有

$$\sum_{\mathbf{Y}} P(Y \mid \pi(Y)) \log P^{\mathscr{G}}(Y \mid \pi(Y)) \geqslant \sum_{\mathbf{Y}} P(Y \mid \pi(Y)) \log Q^{\mathscr{G}}(Y \mid \pi(Y)).$$

所以，

$$\begin{aligned}\sum_{\mathbf{Y}} P(\mathbf{Y}) \log P^{\mathscr{G}}(\mathbf{Y}) \geqslant & \sum_{\mathbf{Y}'} P(\mathbf{Y}') \log Q^{\mathscr{G}}(\mathbf{Y}') \\ & + \sum_{\pi(Y)} P(\pi(Y)) \sum_{Y} P(Y \mid \pi(Y)) \log Q^{\mathscr{G}}(Y \mid \pi(Y)) \\ & = \sum_{\mathbf{Y}} P(\mathbf{Y}) \log Q^{\mathscr{G}}(\mathbf{Y}).\end{aligned}$$

命题得证. □

如果 P 是 $\mathscr{G}$-可分解的，那么 $P^{\mathscr{G}}(\mathbf{X}) = P(\mathbf{X})$. 把这个事实和定理 7.1 放在一起，可得如下结果.

定理 7.3 设数据 $\mathscr{D}$ 是从 $P_{\mathscr{N}}(\mathbf{X})$ 中抽样而得来的. 那么，当样本量 m 趋于无穷时，P^* 几乎必然地收敛到 $P_{\mathscr{N}}$，即

$$P^*(\mathbf{X}) \xrightarrow[m \to \infty]{\text{a. s.}} P_{\mathscr{N}}(\mathbf{X}).$$

□

定理 7.3 表明最大似然估计在样本量趋于无穷时能够找出数据的原生分布，因此它被称为是**一致的**（consistent).

7.6 一般网络贝叶斯估计

在贝叶斯估计的框架中，要对贝叶斯网 $\mathscr{N}=(\mathscr{G}, \boldsymbol{\theta})$ 的参数进行估计，首先要视 $\boldsymbol{\theta}$ 为随机变量，并把关于 $\boldsymbol{\theta}$ 的先验知识表示成一个先验概率分布 $p(\boldsymbol{\theta})$. 要计算的是在观测到 i. i. d. 完整数据 $\mathscr{D}=(\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m)$ 以后，$\boldsymbol{\theta}$ 的后验概率分布 $p(\boldsymbol{\theta} \mid \mathscr{D})$，以及下一个样本 $\mathbf{D}_{m+1}$ 的概率分布 $P(\mathbf{D}_{m+1} \mid \mathscr{D})$.

式 (7.27) 已经给出了 $\boldsymbol{\theta}$ 的对数似然函数与充分统计量 m_{ijk} 的关系. 据此可知，$\boldsymbol{\theta}$ 的似然函数为

$$L(\boldsymbol{\theta} \mid \mathscr{D}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{m_{ijk}}. \tag{7.32}$$

根据贝叶斯公式，有

$$p(\boldsymbol{\theta} \mid \mathscr{D}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{m_{ijk}}. \tag{7.33}$$

为了便于计算，需要对先验概率分布 $p(\boldsymbol{\theta})$ 做一些假设. 在介绍这些假设之前需要先引入两个新的记号. 如前所述，$\boldsymbol{\theta}$ 是所有参数组成的向量，$\boldsymbol{\theta} = \{\theta_{ijk} \mid i=1, \cdots, n; j=1, \cdots, q_i; k=1, \cdots, r_i\}$. 用 $\boldsymbol{\theta}_{ij\cdot}$ 记由 θ_{ij1}，θ_{ij2}，…，θ_{ijr_i} 所组成的子向量，用 $\boldsymbol{\theta}_{i\cdot\cdot}$ 记由 $\boldsymbol{\theta}_{i1\cdot}$，$\boldsymbol{\theta}_{i2\cdot}$，…，$\boldsymbol{\theta}_{iq_i\cdot}$ 所组成的子向量. $\boldsymbol{\theta}_{i\cdot\cdot}$ 包括所有关于

变量 X_i 的条件概率分布 $P(X_i \mid \pi(X_i))$ 的参数，而 $\boldsymbol{\theta}_{ij\cdot}$ 表示所有关于分布 $P(X_i \mid \pi(X_i)=j)$ 的参数.

关于 $p(\boldsymbol{\theta})$ 需要做如下 3 个假设：

假设 7.1 [**全局独立** (global independence) 假设] 关于不同变量 X_i 的参数相互独立，即

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_{i\cdot\cdot}). \tag{7.34}$$

假设 7.2 [**局部独立** (local independence) 假设] 给定一个变量 X_i，对应于 $\pi(X_i)$ 的不同取值的参数相互独立，即

$$p(\boldsymbol{\theta}_{i\cdot\cdot}) = \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij\cdot}). \tag{7.35}$$

假设 7.3 $p(\boldsymbol{\theta}_{ij\cdot})$ 是狄利克雷分布 $D[\alpha_{ij1}, \alpha_{ij1}, \cdots, \alpha_{ijr_i}]$.

在上述 3 个假设下，有

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_{i\cdot\cdot}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij\cdot}) \propto \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1}. \tag{7.36}$$

式 (7.36) 定义的 $p(\boldsymbol{\theta})$ 称为**乘积狄利克雷分布** (product Dirichlet distribution).

把式 (7.36) 代入式 (7.33)，得

$$p(\boldsymbol{\theta} \mid \mathscr{D}) \propto \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1}. \tag{7.37}$$

这即是说，后验分布 $p(\boldsymbol{\theta} \mid \mathscr{D})$ 也是一个乘积狄利克雷分布，$p(\boldsymbol{\theta} \mid \mathscr{D})$ 也具有全局和局部独立性，并且 $p(\boldsymbol{\theta}_{ij\cdot} \mid \mathscr{D})$ 是狄利克雷分布 $D[m_{ij1}+\alpha_{ij1}, m_{ij2}+\alpha_{ij2}, \cdots, m_{ijr_i}+\alpha_{ijr_i}]$.

接下来考虑下一个数据样本 $\mathbf{D}_{m+1}=(X_1, X_2, \cdots, X_n)$ 的概率分布 $P(\mathbf{D}_{m+1} \mid \mathscr{D})$. 首先有

$$P(\mathbf{D}_{m+1} \mid \mathscr{D}) = \int P(\mathbf{D}_{m+1} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathscr{D})\, \mathrm{d}\boldsymbol{\theta}. \tag{7.38}$$

而

$$\begin{aligned} P(\mathbf{D}_{m+1} \mid \boldsymbol{\theta}) &= P(X_1, X_2, \cdots, X_n \mid \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} P(X_i \mid \pi(X_i), \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} P(X_i \mid \pi(X_i), \boldsymbol{\theta}_{i\cdot\cdot}). \end{aligned} \tag{7.39}$$

另一方面，由于 $p(\boldsymbol{\theta} \mid \mathscr{D})$ 具有全局独立性，有

$$p(\boldsymbol{\theta} \mid \mathscr{D}) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_{i\cdot\cdot} \mid \mathscr{D}). \tag{7.40}$$

所以，

$$P(\mathbf{D}_{m+1} \mid \mathscr{D}) = \prod_{i=1}^{n} \int P(X_i \mid \pi(X_i), \boldsymbol{\theta}_{i\cdot\cdot}) p(\boldsymbol{\theta}_{i\cdot\cdot} \mid \mathscr{D}) \, \mathrm{d}\boldsymbol{\theta}_{i\cdot\cdot}$$

$$= \prod_{i=1}^{n} P(X_i \mid \pi(X_i), \mathscr{D}). \tag{7.41}$$

式（7.41）说的是，$P(\mathbf{D}_{m+1} \mid \mathscr{D})$ 是 $\mathscr{G}$-可分解的，因此可以表示为一个以 $\mathscr{G}$ 为结构的贝叶斯网. 用 $\boldsymbol{\theta}'$ 表示这个贝叶斯网的参数：

$$\boldsymbol{\theta}' = \{\theta'_{ijk} \mid i = 1, \cdots, n; j = 1, \cdots, q_i; k = 1, \cdots, r_i\}.$$

注意，θ'_{ijk} 与贝叶斯网 $\mathscr{N}=(\mathscr{G}, \boldsymbol{\theta})$ 的参数 θ_{ijk} 不同：θ'_{ijk} 是一个固定的取值，而 θ_{ijk} 则是一个随机变量. 参数 θ'_{ijk} 可以按如下公式计算：

$$\begin{aligned}\theta'_{ijk} &= P(X_i = k \mid \pi(X_i) = j, \mathscr{D}) \\ &= \int P(X_i = k \mid \pi(X_i) = j, \theta_{ijk}) p(\theta_{ijk} \mid \mathscr{D}) \mathrm{d}\theta_{ijk} \\ &= \int \theta_{ijk} p(\theta_{ijk} \mid \mathscr{D}) \mathrm{d}\theta_{ijk}.\end{aligned}$$

由于 $p(\boldsymbol{\theta}_{ij\cdot})$ 是狄利克雷分布 $D[m_{ij1} + \alpha_{ij1}, m_{ij2} + \alpha_{ij2}, \cdots, m_{ijr_i} + \alpha_{ijr_i}]$，因此有

$$\theta'_{ijk} = \frac{m_{ijk} + \alpha_{ijk}}{\sum_{k=1}^{r_i} (m_{ijk} + \alpha_{ijk})}. \tag{7.42}$$

所以，

$$P(\mathbf{D}_{m+1} | \mathscr{D}) = \prod_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k : \mathbf{D}_{m+1}) \frac{m_{ijk} + \alpha_{ijk}}{\sum_{k=1}^{r_i} (m_{ijk} + \alpha_{ijk})}. \tag{7.43}$$

例 7.7 对图 7.4(a) 所示的贝叶斯网 $\mathscr{N}$ 和图 7.4(b) 所示的完整数据 $\mathscr{D}$，考虑计算 $\mathscr{N}$ 参数的贝叶斯估计. 首先假设先验分布 $p(\boldsymbol{\theta})$ 是乘积狄利克雷分布，且其超参数 $\boldsymbol{\alpha}^0 = \{\alpha^0_{ijk} \mid i=1,2,3; j=1,\cdots,q_i; k=1,2\}$ 分别如下：

α^0_{1jk}

j \ k	1	2
1	2	2

α^0_{2jk}

j \ k	1	2
1	1	1
2	1	1

α^0_{3jk}

j \ k	1	2
1	1	1
2	1	1

这等于是假设关于 $\boldsymbol{\theta}$ 的先验知识相当于如下的虚拟数据 $\mathscr{D}'$：

	X_1	X_2	X_3
$\mathscr{D}'_1$	1	1	1
$\mathscr{D}'_2$	1	2	1
$\mathscr{D}'_3$	2	1	2
$\mathscr{D}'_4$	2	2	2

它的等价样本量为 4. 根据式（7.37），后验分布 $p(\boldsymbol{\theta} \mid \mathscr{D})$ 也是乘积狄利克雷分布，其超参数 $\boldsymbol{\alpha}$ 为

α_{1jk}

j \ k	1	2
1	4	4

α_{2jk}

j \ k	1	2
1	3	1
2	1	3

α_{3jk}

j \ k	1	2
1	2	2
2	1	3

下一个样本的分布 $P(\mathbf{D}_{m+1} \mid \mathscr{D})$ 可以表示为结构如 7.4（a）所示的贝叶斯网. 根据式（7.42），其参数 $\boldsymbol{\theta}$ 为

$P(X_1)$

X_1	1	2
	4/8	4/8

$P(X_2 \mid X_1)$

X_1 \ X_2	1	2
1	3/4	1/4
2	1/4	3/4

$P(X_3 \mid X_2)$

X_2 \ X_3	1	2
1	2/4	2/4
2	1/4	3/4

□

7.7　缺值数据最大似然估计

本节和 7.8 节分别讨论如何在缺值数据情况下进行最大似然估计和贝叶斯估计. 在分析缺值数据时，人们往往假设数据是**随机缺失**（missing at random, MAR）的，即一个变量值的缺失与它的实际取值无关，从而不能基于前者对后者做任何推测. 设 $\mathbf{D}$ 是一个样本，$\mathbf{O}$ 是所有其值在 $\mathbf{D}$ 中已知的变量的集合，H 是其值在 $\mathbf{D}$ 未知的一个变量. 随机缺失假设可表述为

$$P(H \mid H\text{- 值缺}, \mathbf{O}) = P(H \mid \mathbf{O}). \tag{7.44}$$

随机缺失假设有时不成立. 例如，一个病人的病历中若没有“X-光检测结果”的数据，往往是因为医生认为没有必要进行该项检验，如果检验了，结果应该是阴性. 当式（7.44）不成立时，可以引入一个辅助变量 H_{obs}，当 H 的取值被观测到时，它的取值为“真”；而当 H 值缺时，它的取值为“假”. 显然 H_{obs} 的取值总是已知的，而且

$$P(H \mid H\text{- 值缺}, H_{\text{obs}}, \mathbf{O}) = P(H \mid H_{\text{obs}}, \mathbf{O}). \tag{7.45}$$

所以通过引入辅助变量 H_{obs}，总可以保证随机缺失假设成立. 注意，引入 H_{obs} 意味着在建立贝叶斯网的结构时需要把它考虑进去.

在完整数据情况下，最大似然估计可以用一个**闭公式**（closed formula），即式（7.28），来计算. 但当数据有缺值时，$\log P(\mathbf{D}_l \mid \boldsymbol{\theta})$ 不能按式（7.25）的形式分解，对数似然函数 $l(\boldsymbol{\theta} \mid \mathscr{D})$ 也不能写成式（7.27）的形式，因而没有计算最

大似然估计的闭公式. 在实际中，人们通常使用迭代法对它进行近似. 下面介绍最常用的迭代算法，即**期望优化**（expectation maximization）算法，简称**EM算法**. 我们先介绍EM算法的基本思想和基本理论，然后讨论算法的实现及其收敛性.

7.7.1 EM算法的基本思想

为了求得贝叶斯网 $\mathcal{N}=(\mathcal{G}, \boldsymbol{\theta})$ 的参数 $\boldsymbol{\theta}$ 之MLE的近似值，EM算法从 $\boldsymbol{\theta}$ 的某个初始值 $\boldsymbol{\theta}^0$ 出发开始迭代，初始值 $\boldsymbol{\theta}^0$ 往往是随机产生的. 设已经进行了 t 次迭代，得到估计 $\boldsymbol{\theta}^t$. 第 $t+1$ 次迭代由以下两个步骤组成：

(1) 基于 $\boldsymbol{\theta}^t$ 对数据进行修补，使之完整；

(2) 基于修补后的完整数据计算 $\boldsymbol{\theta}$ 的最大似然估计，得 $\boldsymbol{\theta}^{t+1}$.

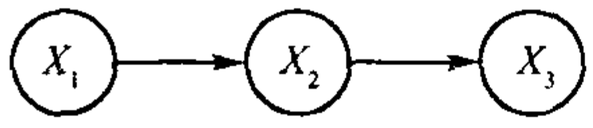
下面逐一描述这两个步骤. 考虑某一缺值样本 $\mathbf{D}_l$，设 $\mathbf{X}_l$ 为 $\mathbf{D}_l$ 中所有值缺变量的集合. 对 $\mathbf{X}_l$ 的一个取值 $\mathbf{x}_l$，将 $\mathbf{X}_l=\mathbf{x}_l$ 加入 $\mathbf{D}_l$，就得到一完整样本，$(\mathbf{D}_l, \mathbf{X}_l=\mathbf{x}_l)$，这个过程称为**数据修补**（data completion）. 由于 $\mathbf{X}_l$ 有多个可能取值，所以 $\mathbf{D}_l$ 有多种修补方式. EM算法考虑所有可能的修补结果，并给每一结果附加一个权重 $w_{\mathbf{x}_l}$，得到加权样本 $(\mathbf{D}_l, \mathbf{X}_l=\mathbf{x}_l)[w_{\mathbf{x}_l}]$. 为了确定这个权重，EM算法考虑当前估计 $\boldsymbol{\theta}^t$ 与网络结构 $\mathcal{G}$ 组成的贝叶斯网 $(\mathcal{G}, \boldsymbol{\theta}^t)$，在其中计算 $P(\mathbf{X}_l=\mathbf{x}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t)$，所得的结果为样本 $(\mathbf{D}_l, \mathbf{X}_l=\mathbf{x}_l)$ 的权重. 这个权重在0和1之间，因此加权样本 $(\mathbf{D}_l, \mathbf{X}_l=\mathbf{x}_l)[w_{\mathbf{x}_l}]$ 又称为**碎权样本**（fractional sample）.

在数据修补过程中，每个缺值样本都被一系列完整的碎权样本所代替. 所以，在修补后的数据中所有样本都是完整的，而且每个样本都有一个权重，原本完整的样本权重为1. 与式（7.28）类似，EM算法用下式计算基于补后数据的最大似然估计 $\boldsymbol{\theta}^{t+1}$，即

$$\theta_{ijk}^{t+1}=\begin{cases}\dfrac{m_{ijk}^t}{\sum_{k=1}^{r_i} m_{ijk}^t}, & \text{若} \sum_{k=1}^{r_i} m_{ijk}^t>0 \\ \dfrac{1}{r_i}, & \text{若否}\end{cases} \tag{7.46}$$

其中 m_{ijk}^t 是补后数据中所有满足 $X_i=k$ 和 $\pi(X_i)=j$ 的样本的权重之和.

例7.8 设图7.5（a）是一个贝叶斯网 $\mathcal{N}$，其中所有变量均取二值，1或2. 设图7.5（b）所示是从 $\mathcal{N}$ 的联合分布中抽样而得到的一组数据，它由两个完整样本 $\mathbf{D}_1$，$\mathbf{D}_2$ 和两个缺值样本 $\mathbf{D}_3$，$\mathbf{D}_4$ 组成. 考虑用EM算法来近似计算 $\mathcal{N}$ 参数的MLE估计.



(a) 贝叶斯网 $\mathcal{N}$

	X_1	X_2	X_3
$\mathbf{D}_1$	1	1	1
$\mathbf{D}_2$	2	2	2
$\mathbf{D}_3$	1	—	1
$\mathbf{D}_4$	2	—	2

(b) 缺值数据

图 7.5　一个贝叶斯网 $\mathcal{N}$ 以及一组缺值数据

EM 首先选择一组初始参数值 $\boldsymbol{\theta}^0$. 设 $\boldsymbol{\theta}^0$ 如下所示：

$P(X_1)$

X_1	1	2
	1/2	1/2

$P(X_2 \mid X_1)$

X_1 \ X_2	1	2
1	2/3	1/3
2	1/3	2/3

$P(X_3 \mid X_2)$

X_2 \ X_3	1	2
1	2/3	1/3
2	1/3	2/3

从 $\boldsymbol{\theta}^0$ 出发，EM 开始迭代. 在第一次迭代中，先用 $\boldsymbol{\theta}^0$ 修补数据. 由于

$$P(X_2 = 1 \mid \mathbf{D}_3, \boldsymbol{\theta}^0) = 4/5, \qquad P(X_2 = 2 \mid \mathbf{D}_3, \boldsymbol{\theta}^0) = 1/5,$$

所以 $\mathbf{D}_3$ 被如下两个碎权样本所替换：$\mathbf{D}_{3.1} = (1, 1, 1)\left[\frac{4}{5}\right]$，$\mathbf{D}_{3.2} = (1, 2, 1)\left[\frac{1}{5}\right]$. 类似地，$\mathbf{D}_4$ 也被两个碎权样本替换. 修补后得到如下的碎权完整数据：

	X_1	X_2	X_3	权重
$\mathbf{D}_1$	1	1	1	1
$\mathbf{D}_2$	2	2	2	1
$\mathbf{D}_{3.1}$	1	1	1	4/5
$\mathbf{D}_{3.2}$	1	2	1	1/5
$\mathbf{D}_{4.1}$	2	1	1	1/5
$\mathbf{D}_{4.2}$	2	2	2	4/5

根据式 (7.46)，第一次迭代完成后的估计 $\boldsymbol{\theta}^1$ 如下：

$P(X_1)$

X_1	1	2
	1/2	1/2

$P(X_2 \mid X_1)$

X_1 \ X_2	1	2
1	9/10	1/10
2	1/10	9/10

$P(X_3 \mid X_2)$

X_2 \ X_3	1	2
1	9/10	1/10
2	1/10	9/10

EM 接着进入第二次迭代. 先用 $\boldsymbol{\theta}^1$ 对数据进行修补. 由于

$$P(X_2=1 \mid \mathbf{D}_3, \boldsymbol{\theta}^1)=81/82, P(X_2=2 \mid \mathbf{D}_3, \boldsymbol{\theta}^1)=1/82,$$

$$P(X_2=2 \mid \mathbf{D}_4, \boldsymbol{\theta}^1)=1/82, P(X_2=2 \mid \mathbf{D}_4, \boldsymbol{\theta}^1)=81/82,$$

修补后的数据如下：

	X_1	X_2	X_3	权重
$\mathbf{D}_1$	1	1	1	1
$\mathbf{D}_2$	2	2	2	1
$\mathbf{D}_{3.1}$	1	1	1	81/82
$\mathbf{D}_{3.2}$	1	2	1	1/82
$\mathbf{D}_{4.1}$	2	1	1	1/82
$\mathbf{D}_{4.2}$	2	2	2	81/82

于是，在第二次迭代完成后得到的估计 $\boldsymbol{\theta}^2$ 如下：

$P(X_1)$

X_1	1	2
	1/2	1/2

$P(X_2 \mid X_1)$

X_1 \ X_2	1	2
1	$\frac{163}{164}$	$\frac{1}{164}$
2	$\frac{1}{164}$	$\frac{163}{164}$

$P(X_3 \mid X_2)$

X_2 \ X_3	1	2
1	$\frac{163}{164}$	$\frac{1}{164}$
2	$\frac{1}{164}$	$\frac{163}{164}$

如此迭代下去. 容易看到，这个过程快速地向如下结果收敛：

$P(X_1)$

X_1	1	2
	1/2	1/2

$P(X_2 \mid X_1)$

X_1 \ X_2	1	2
1	1	0
2	0	1

$P(X_3 \mid X_2)$

X_2 \ X_3	1	2
1	1	0
2	0	1

□

7.7.2 EM 算法的基本理论

数据修补的思想直观易懂，但是为什么式（7.46）给出的就是基于修补后的碎权完整数据的最大似然估计呢？为了回答这个问题，首先需要将似然函数的概念推广到补后数据.

设 $\mathscr{D}=(\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m)$ 是关于某贝叶斯网 $\mathscr{N}=(\mathscr{G}, \boldsymbol{\theta})$ 的一组 i. i. d. 数据. 对其中任一样本 $\mathbf{D}_l$，设 $\mathbf{X}_l$ 是 $\mathbf{D}_l$ 中所有值缺变量的集合. 设 $\boldsymbol{\theta}^t$ 是关于参数 $\boldsymbol{\theta}$ 的当前估计，$\mathscr{D}^t$ 是基于 $\boldsymbol{\theta}^t$ 将 $\mathscr{D}$ 修补而得到的碎权完整数据. 定义 $\boldsymbol{\theta}$ 的基于 $\mathscr{D}^t$ 的对数似然函数为

$$l(\boldsymbol{\theta} \mid \mathscr{D}^t) = \sum_{l=1}^{m} \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l = \mathbf{x}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l \mid \boldsymbol{\theta}). \quad (7.47)$$

其中 $P(\mathbf{X}_l = \mathbf{x}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t)$ 就是 $w_{\mathbf{x}_l}$. 当 $\mathbf{X}_l = \varnothing$ 时,约定 $P(\mathbf{X}_l = \mathbf{x}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t)$ 为 1.

由于 $\mathscr{D}^t$ 完全由 $\mathscr{D}$ 和 $\boldsymbol{\theta}^t$ 决定,$l(\boldsymbol{\theta} \mid \mathscr{D}^t)$ 一般写成 $l(\boldsymbol{\theta} \mid \mathscr{D}, \boldsymbol{\theta}^t)$,并被称为是 $\boldsymbol{\theta}$ 的基于 $\mathscr{D}$ 的**期望对数似然函数** (expected loglikelihood function). 在 EM 算法的迭代过程中,数据 $\mathscr{D}$ 是不变的,因此 $l(\boldsymbol{\theta} \mid \mathscr{D}, \boldsymbol{\theta}^t)$ 往往被简记为 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)$.

在 EM 的第 t 次迭代过程中,第一步计算期望对数似然函数 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)$,因此称为 **E-步骤** (E-step)①. 第二步求得使 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)$ 达到最大的 $\boldsymbol{\theta}$ 的取值,即

$$\boldsymbol{\theta}^{t+1} = \arg\sup_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t), \quad (7.48)$$

因此称为 **M-步骤** (M-step)② .

下面讨论如何计算 $\boldsymbol{\theta}^{t+1}$. 首先,碎权样本 $(\mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l)$ 的特征函数为

$$\chi(i, j, k; \mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l) = \begin{cases} 1, & \text{若在}(\mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l)\text{中 } X_i = k \text{ 且 } \pi(X_i) = j \\ 0, & \text{若否} \end{cases}. \quad (7.49)$$

由于式 (7.25),有

$$\log P(\mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k; \mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l) \log \theta_{ijk}.$$

于是,

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t) &= \sum_{l=1}^{m} \sum_{\mathbf{x}_l \in \Omega_{\mathbf{x}_l}} P(\mathbf{X}_l = \mathbf{x}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k; \mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l) \log \theta_{ijk} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \sum_{l=1}^{m} \sum_{\mathbf{x}_l \in \Omega_{\mathbf{x}_l}} P(\mathbf{X}_l = \mathbf{x}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \chi(i, j, k; \mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l) \log \theta_{ijk}. \end{aligned}$$

定义

$$m_{ijk}^t = \sum_{l=1}^{m} \sum_{\mathbf{x}_l \in \Omega_{\mathbf{x}_l}} P(\mathbf{X}_l = \mathbf{x}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \chi(i, j, k; \mathbf{D}_l, \mathbf{X}_l = \mathbf{x}_l). \quad (7.50)$$

直观上,m_{ijk}^t 是补后数据 $\mathscr{D}^t$ 中所有满足 $X_i = k$ 和 $\pi(X_i) = j$ 的样本的权重之和. 进而有

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk}^t \log \theta_{ijk}. \quad (7.51)$$

根据推论 1.1,当 $\boldsymbol{\theta}$ 取如下值时,$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)$ 达到最大:

---

① E 是 expectation 的缩写.

② M 是 maximization 的缩写.

$$\theta_{ijk}^{t+1}=\begin{cases}\dfrac{m_{ijk}^{t}}{\sum_{k=1}^{r_i} m_{ijk}^{t}}, & 若\sum_{k=1}^{r_i} m_{ijk}^{t}>0\\ \dfrac{1}{r_i}, & 若否\end{cases}. \tag{7.52}$$

这样我们就得到了式（7.46）.

7.7.3　EM 算法

EM 的 E-步骤要计算的是 $Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^t)$．根据式（7.51），这就是要计算充分统计量 m_{ijk}^t．不难看出，

$$\sum_{\mathbf{x}_l\in\Omega_{\mathbf{X}_l}} P(\mathbf{X}_l=\mathbf{x}_l\mid\mathbf{D}_l,\boldsymbol{\theta}^t)\chi(i,j,k:\mathbf{D}_l,\mathbf{X}_l=\mathbf{x}_l)=P(X_i=k,\pi(X_i)=j\mid\mathbf{D}_l,\boldsymbol{\theta}^t).$$

所以，m_{ijk}^t 可以用下式来计算：

$$m_{ijk}^t=\sum_{l=1}^{m}P(X_i=k,\pi(X_i)=j\mid\mathbf{D}_l,\boldsymbol{\theta}^t). \tag{7.53}$$

EM 的 M-步骤要计算 $\arg\sup_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^t)$，这只需把从式（7.53）得到的 m_{ijk}^t 代入式（7.52）即可．EM 算法的伪代码由图 7.6 给出，其中第 7 步的收敛条件将会在 7.7.4 小节讨论.

EM($\mathscr{G},\mathscr{D},\delta$)

输入：$\mathscr{G}$——贝叶斯网 $\mathscr{N}$ 的结构；
　　$\mathscr{D}$——一组关于 $\mathscr{N}$ 中变量的数据；　δ——收敛阈值.
输出：$\mathscr{N}$ 的参数的估计；

```
1: t←0, θ^0←随机参数值;
2: oldScore←l(θ^t|𝒟);
3: while (true)
4:     E-步骤:按式(7.53)计算 m_ijk^t (i=1,…,n; j=1,…,q_i; k=1,…,r_i);
5:     M-步骤:按式(7.52)计算 θ^{t+1};
6:     newScore←l(θ^{t+1}|𝒟);
7:     if (newScore>oldScore+δ)
8:         oldScore←newScore;
9:         t←t+1;
10:    else
11:        return θ^{t+1}
12:    end if
13: end while.
```

图 7.6　EM 算法

下面简单谈一谈 EM 算法实现的问题. 首先是参数向量 $\boldsymbol{\theta}^t$ 该如何表示. 一种做法是把它表示为一个一维数组，但这使用起来不方便. 因为 $\boldsymbol{\theta}^t$ 是贝叶斯网中的所有参数所组成的向量，所以一个比较好的办法是用贝叶斯网的数据结构来表示它.

假设已知 $\boldsymbol{\theta}^t$，怎样来计算 $\boldsymbol{\theta}^{t+1}$ 呢? 一个直接的方法是对每一个 i，先按式 (7.53) 调用变量消元算法 m 次来计算 $m_{ijk}^t(j=1,\cdots,q_i;k=1,\cdots,r_i)$，然后再按式 (7.52) 计算相应的 θ_{ijk}^{t+1}. 这样，在计算式 (7.53) 时会进行 mn 次推理，在各次推理之间没有实现计算共享. 下面介绍一种基于团树传播，实现计算共享，从而加快 $\boldsymbol{\theta}^{t+1}$ 的计算的方法.

设 $\mathscr{T}$ 是一棵覆盖 $\mathscr{N}$ 的团树. 这个方法由以下 4 步组成:

(1) 将 $m_{ijk}^t(j=1,\cdots,q_i;k=1,\cdots,r_i)$ 置零;

(2) 按照第 5.2 节所述的方法，用 $\boldsymbol{\theta}^t$ 将团树 $\mathscr{T}$ 初始化;

(3) 对每一个样本 $\mathbf{D}_l$:

①在团树 $\mathscr{T}$ 中按 $\mathbf{D}_l$ 设置证据，并调用算法 CTP 的第 3～5 步进行信息传递;

②对每一个变量 X_i:

- 在团树 $\mathscr{T}$ 中找到 X_i 的一个家族覆盖团 $\mathbf{C}_{X_i}$，调用算法 CTP 的第 6.2 步，得到一函数 $h(\mathbf{C}_{X_i})$，进而得到①

$$P(X_i,\pi(X_i)\mid \mathbf{D}_l,\boldsymbol{\theta}^t)=\sum_{\mathbf{C}_{X_i}\setminus\{X_i\}\cup\pi(X_i)} h(\mathbf{C}_{X_i}); \tag{7.54}$$

- 在 $m_{ijk}^t(j=1,\cdots,q_i;k=1,\cdots,r_i)$ 中加上式 (7.54) 所得的结果;

(4) 将 $m_{ijk}^t(i=1,\cdots,n;j=1,\cdots,q_i;k=1,\cdots,r_i)$ 代入式 (7.52)，从而获得 $\boldsymbol{\theta}^{t+1}$.

接下来讨论 EM 算法的第 6 步如何计算 $l(\boldsymbol{\theta}^{t+1}\mid\mathscr{D})$. 由于 i. i. d. 假设，可以对每一个样本 $\mathbf{D}_l$ 分别调用 VE 算法来计算 $l(\boldsymbol{\theta}^{t+1}\mid\mathbf{D}_l)$，它们的和就是对数似然度 $l(\boldsymbol{\theta}^{t+1}\mid\mathscr{D})$. 要指出的是，将上述计算 $\boldsymbol{\theta}^{t+1}$ 的方法稍加修改，就可以在计算 $\boldsymbol{\theta}^{t+1}$ 的同时获得对数似然度 $l(\boldsymbol{\theta}^{t+1}\mid\mathscr{D})$，从而实现计算共享. 其基本思想是在信息传递完成之后，任选一个团 $\mathbf{C}$，将传递到 $\mathbf{C}$ 的信息以及存储在 $\mathbf{C}$ 处的函数相乘，得到一个函数 $h(\mathbf{C})$. 由命题 5.1 知，$\sum_{\mathbf{C}} h(\mathbf{C})$ 就是 $l(\boldsymbol{\theta}^{t+1}\mid\mathbf{D}_l)$.

7.7.4 EM 算法的收敛性

现在来看看 EM 算法在运行过程中产生的序列 $\{l(\boldsymbol{\theta}_0\mid\mathscr{D}),l(\boldsymbol{\theta}_1\mid\mathscr{D}),l(\boldsymbol{\theta}_2\mid\mathscr{D}),\cdots\}$ 的性质.

---

① 注意，这里忽略了一个细节，即 $\mathbf{C}_{X_i}$ 中包含证据变量的情况.

定理 7.4 设$\boldsymbol{\theta}^t$和$\boldsymbol{\theta}^{t+1}$分别是EM算法在第$t$次和$t+1$次迭代中所获得的结果. 那么有

$$l(\boldsymbol{\theta}^t \mid \mathscr{D}) \leqslant l(\boldsymbol{\theta}^{t+1} \mid \mathscr{D}). \tag{7.55}$$

证明 由于

$$\sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}) = 1, P(\mathbf{D}_l \mid \boldsymbol{\theta}) = \frac{P(\mathbf{D}_l, \mathbf{X}_l \mid \boldsymbol{\theta})}{P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta})},$$

有

$$\begin{aligned}
l(\boldsymbol{\theta} \mid \mathscr{D}) &= \sum_{l=1}^{m} \log P(\mathbf{D}_l \mid \boldsymbol{\theta}) \\
&= \sum_{l=1}^{m} \sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log \frac{P(\mathbf{D}_l, \mathbf{X}_l \mid \boldsymbol{\theta})}{P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta})} \\
&= \sum_{l=1}^{m} \sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{D}_l, \mathbf{X}_l \mid \boldsymbol{\theta}) \\
&\quad - \sum_{l=1}^{m} \sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}) \\
&= Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t) - \sum_{l=1}^{m} \sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}).
\end{aligned}$$

根据式(7.48), 有$Q(\boldsymbol{\theta}^t \mid \boldsymbol{\theta}^t) \leqslant Q(\boldsymbol{\theta}^{t+1} \mid \boldsymbol{\theta}^t)$. 而根据信息不等式(定理1.3), 有

$$\sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \geqslant \sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^{t+1}).$$

所以,

$$\begin{aligned}
l(\boldsymbol{\theta}^t \mid \mathscr{D}) &= Q(\boldsymbol{\theta}^t \mid \boldsymbol{\theta}^t) - \sum_{l=1}^{m} \sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \\
&\leqslant Q(\boldsymbol{\theta}^{t+1} \mid \boldsymbol{\theta}^t) - \sum_{l=1}^{m} \sum_{\mathbf{X}_l} P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^t) \log P(\mathbf{X}_l \mid \mathbf{D}_l, \boldsymbol{\theta}^{t+1}) \\
&= l(\boldsymbol{\theta}^{t+1} \mid \mathscr{D}).
\end{aligned}$$

定理得证. □

推论 7.1 设$\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \cdots\}$是EM算法所获得的参数估计序列. 则序列$\{l(\boldsymbol{\theta}^0 \mid \mathscr{D}), l(\boldsymbol{\theta}^1 \mid \mathscr{D}), l(\boldsymbol{\theta}^2 \mid \mathscr{D}), \cdots\}$是单调递增的, 并且收敛.

证明 定理7.4已证明了序列是单调递增的. 另一方面, 它又是有上界的, 0就是它的一个上界. 所以它必然收敛. 推论得证. □

EM算法依据两次相邻迭代所得估计的似然度之差$\delta_t = l(\boldsymbol{\theta}^{t+1} \mid \mathscr{D}) - l(\boldsymbol{\theta}^t \mid \mathscr{D})$来决定是否继续进行迭代. 若$\delta_t$小于某个预先设定的阈值$\delta$, EM就停止迭代, 返回

$\boldsymbol{\theta}^{t+1}$，否则就继续. 根据推论 7.1，只要 $\delta > 0$，当 t 足够大时，δ_t 一定会小于 δ，所以 EM 算法最终一定会停止迭代，返回估计.

EM 算法虽然收敛，但不一定收敛于全局最优，也可能收敛于局部最优或鞍点. 因此，它给出的估计可能与最大似然估计相差甚远. 为了提高估计的质量，人们往往从不同的初始点出发多次运行 EM 算法，然后把所得结果进行比较，选择似然度最大的那个估计作为最后的结果.

EM 算法的收敛速度开始时比较快，后来渐渐放慢. 图 7.7 给出了一个具有代表性的例子. EM 算法的收敛速度与数据缺失的多少有关. 一般来讲，数据缺失越多，收敛速度就越慢. 详细情况参见文献（McLachlan and Krishnan，1997）.

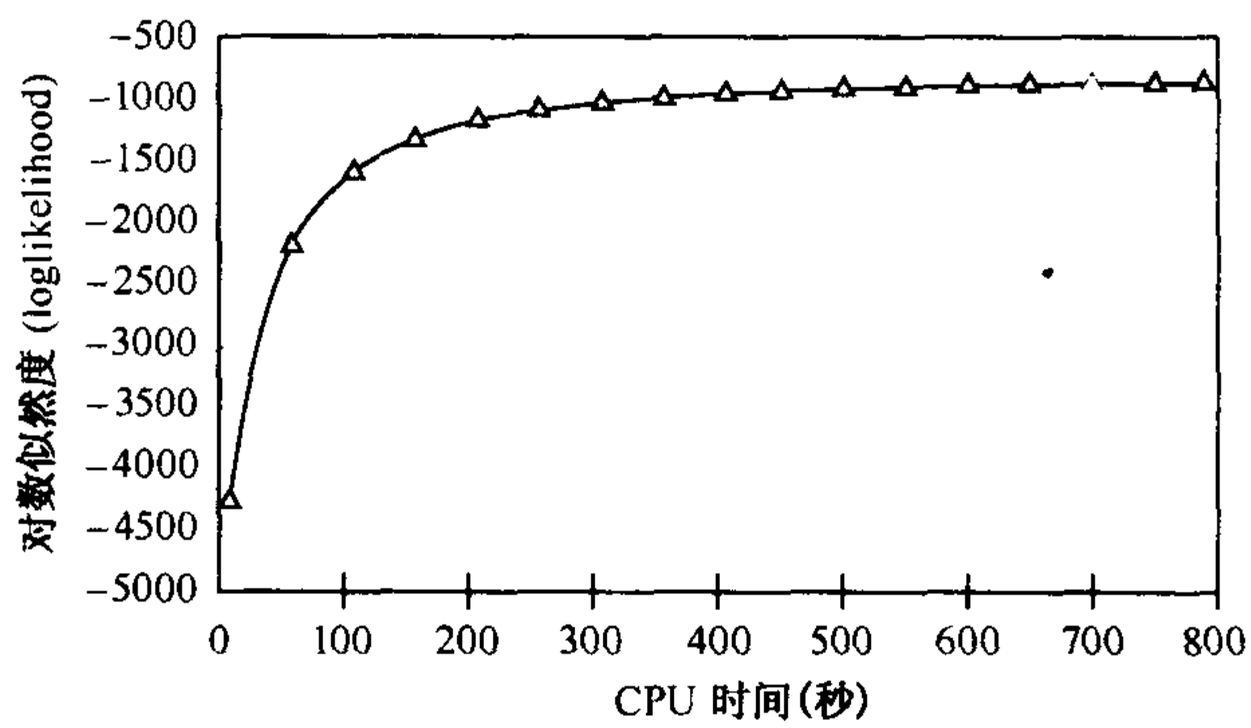


图 7.7 EM 算法的收敛速度

7.8 缺值数据贝叶斯估计

下面回到对贝叶斯网 $\mathcal{N}=(\mathcal{G}, \boldsymbol{\theta})$ 的参数做贝叶斯估计的问题. 当数据完整时，贝叶斯估计可以用式（7.37）和式（7.42）来计算. 当数据有缺值时，似然函数不能写成式（7.32）的形式，因而没有计算贝叶斯估计的闭公式. 在实际中，人们往往采用近似方法. 下面介绍一种最简单的近似方法，即**碎权更新** (fractional updating).

碎权更新首先设先验概率分布 $p(\boldsymbol{\theta})$ 是乘积狄利克雷分布，然后按一定顺序逐个处理数据样本. 在处理下一个样本时，它先利用当前估计对样本进行修补，得到一组碎权完整样本，然后用这些样本将 $\boldsymbol{\theta}$ 的估计更新. 在整个过程中，对 $\boldsymbol{\theta}$ 的估计不断变化，但始终是乘积狄利克雷分布.

假设已处理完样本 $\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l$，得到 $p(\boldsymbol{\theta} \mid \mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l)$ 的一个近似，它是一个乘积狄利克雷分布，参数为 $\boldsymbol{\alpha}^l = \{\alpha_{ijk}^l \mid i=1, \cdots, n; j=1, \cdots, q_i; k=1, \cdots, r_i\}$. 基于这个估计，先计算下一个样本 $\mathbf{D}_{l+1}$ 的分布 $P(\mathbf{D}_{l+1} \mid \mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l)$. 在第 7.6

节已看到，$P(\mathbf{D}_{l+1} \mid \mathbf{D}_1, \cdots, \mathbf{D}_l)$可以用贝叶斯网 $\mathcal{N}^l=(\mathcal{G}, \boldsymbol{\theta}^l)$来表示，其中，

$$\theta_{ijk}^l = \frac{\alpha_{ijk}^l}{\sum_{k=1}^{r_i} \alpha_{ijk}^l}. \tag{7.56}$$

注意，$\mathcal{N}^l$的参数θ_{ijk}^l是固定的取值，而$\mathcal{N}$的参数θ_{ijk}是随机变量. 为方便起见，下面将分布 $P(\mathbf{D}_{l+1} \mid \mathbf{D}_1, \cdots, \mathbf{D}_l)$ 记为 P^l.

设$\mathbf{X}_{l+1}$为$\mathbf{D}_{l+1}$中所有缺值变量的集合. 考虑$\mathbf{X}_{l+1}$的某一取值$\mathbf{x}_{l+1}$. 根据当前的估计，$\mathbf{X}_{l+1}=\mathbf{x}_{l+1}$的概率是 $P^l(\mathbf{X}_{l+1}=\mathbf{x}_{l+1})$. 于是，我们对$\mathbf{D}_{l+1}$进行修补，得碎权完整数据 $(\mathbf{D}_{l+1}, \mathbf{X}_{l+1}=\mathbf{x}_{l+1})$ $[P^l(\mathbf{X}_{l+1}=\mathbf{x}_{l+1})]$. 用这样的补后数据通过式(7.42)将 $p(\boldsymbol{\theta} \mid \mathbf{D}_1, \cdots, \mathbf{D}_l)$ 更新，得到另一个乘积狄利克雷分布，其参数α^{l+1} 为

$$\alpha_{ijk}^{l+1} = \alpha_{ijk}^l + P(X_i = k, \pi(X_i) = j \mid \mathbf{D}_1, \cdots, \mathbf{D}_{l+1}). \tag{7.57}$$

例 7.9 对图 7.5 (a) 和 7.5 (b) 所示的贝叶斯网$\mathcal{N}$和数据$\mathcal{D}$，用碎权更新计算$\mathcal{N}$的参数的贝叶斯估计. 和例 7.7 一样，首先假设先验分布 $p(\boldsymbol{\theta})$ 是乘积狄利克雷分布，其超参数$\boldsymbol{\alpha}^0$ 如下：

α_{1jk}^0

$j \backslash k$	1	2
1	2	2

α_{2jk}^0

$j \backslash k$	1	2
1	1	1
2	1	1

α_{3jk}^0

$j \backslash k$	1	2
1	1	1
2	1	1

从$\boldsymbol{\alpha}^0$ 出发，首先考虑第一个样本$\mathbf{D}_1=(1,1,1)$. 它是完整的，不需要修补. 用$\mathbf{D}_1$ 更新 $p(\boldsymbol{\theta})$，得到 $p(\boldsymbol{\theta} \mid \mathbf{D}_1)$. 它也是一个乘积狄利克雷分布，其超参数$\boldsymbol{\alpha}^1$ 如下：

α_{1jk}^1

$j \backslash k$	1	2
1	3	2

α_{2jk}^1

$j \backslash k$	1	2
1	2	1
2	1	1

α_{3jk}^1

$j \backslash k$	1	2
1	2	1
2	1	1

接下来考虑 $\mathbf{D}_2=(2, 2, 2)$. 它也是完整的，也不需要修补. 用它更新 $p(\boldsymbol{\theta} \mid \mathbf{D}_1)$，得到 $p(\boldsymbol{\theta} \mid \mathbf{D}_1, \mathbf{D}_2)$，它也是一个乘积狄利克雷分布，其超参数$\boldsymbol{\alpha}^2$ 如下：

α_{1jk}^2

$j \backslash k$	1	2
1	3	3

α_{2jk}^2

$j \backslash k$	1	2
1	2	1
2	1	2

α_{3jk}^2

$j \backslash k$	1	2
1	2	1
2	1	2

接下来考虑 $\mathbf{D}_3=(1,-,1)$，其中 X_2 值缺，需要修补．为此考虑 $P(\mathbf{D}_3 \mid \mathbf{D}_1, \mathbf{D}_2)$，它可表示成一个贝叶斯网 $(\mathscr{G}, \boldsymbol{\theta}^2)$，其参数如下：

$P(X_1 \mid \boldsymbol{\theta}^2)$

X_1	1	2
	3/6	3/6

$P(X_2 \mid X_1, \boldsymbol{\theta}^2)$

X_1 \ X_2	1	2
1	$\frac{2}{3}$	$\frac{1}{3}$
2	$\frac{1}{3}$	$\frac{2}{3}$

$P(X_3 \mid X_2, \boldsymbol{\theta}^2)$

X_2 \ X_3	1	2
1	$\frac{2}{3}$	$\frac{1}{3}$
2	$\frac{1}{3}$	$\frac{2}{3}$

易见 $P(X_2=1 \mid \mathbf{D}_3, \boldsymbol{\theta}^2)=\frac{4}{5}$，$P(X_2=2 \mid \mathbf{D}_3, \boldsymbol{\theta}^2)=\frac{1}{5}$．所以得到如下碎权完整样本：$\mathbf{D}_{3.1}=(1,1,1)\left[\frac{4}{5}\right]$，$\mathbf{D}_{3.2}=(1,2,1)\left[\frac{1}{5}\right]$．用这两个样本更新 $p(\boldsymbol{\theta} \mid \mathbf{D}_1, \mathbf{D}_2)$，得到 $p(\boldsymbol{\theta} \mid \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$，它也是乘积狄利克雷分布，其超参数 $\boldsymbol{\alpha}^3$ 如下：

α_{1jk}^3

j \ k	1	2
1	4	3

α_{2jk}^3

j \ k	1	2
1	$\frac{14}{5}$	$\frac{6}{5}$
2	1	2

α_{3jk}^3

j \ k	1	2
1	$\frac{14}{5}$	1
2	$\frac{6}{5}$	2

最后考虑 $\mathbf{D}_4=(2,-,2)$，其中 X_2 值缺，需要修补．为此，考虑 $P(\mathbf{D}_4 \mid \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$，它可以表示成一个贝叶斯网 $(\mathscr{G}, \boldsymbol{\theta}^3)$，其参数如下：

$P(X_1 \mid \boldsymbol{\theta}^3)$

X_1	1	2
	4/7	3/7

$P(X_2 \mid X_1, \boldsymbol{\theta}^3)$

X_1 \ X_2	1	2
1	$\frac{7}{10}$	$\frac{3}{10}$
2	$\frac{1}{3}$	$\frac{2}{3}$

$P(X_3 \mid X_2, \boldsymbol{\theta}^3)$

X_2 \ X_3	1	2
1	$\frac{14}{19}$	$\frac{5}{19}$
2	$\frac{3}{8}$	$\frac{5}{8}$

易见 $P(X_2=1 \mid \mathbf{D}_4, \boldsymbol{\theta}^3)=\frac{4}{23}$，$P(X_2=2 \mid \mathbf{D}_4, \boldsymbol{\theta}^3)=\frac{19}{23}$．所以得到如下碎权完整样本：$\mathbf{D}_{4.1}=(2,1,2)\left[\frac{4}{23}\right]$，$\mathbf{D}_{4.2}=(2,2,2)\left[\frac{19}{23}\right]$．用这两个样本更新 $p(\boldsymbol{\theta} \mid \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$，得到 $p(\boldsymbol{\theta} \mid \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3, \mathbf{D}_4)$．它也是乘积狄利克雷分布，其超参数 $\boldsymbol{\alpha}^4$ 如下：

$\overset{4}{\alpha}_{1jk}$

j \ k	1	2
1	4	4

$\overset{4}{\alpha}_{2jk}$

j \ k	1	2
1	$\frac{14}{5}$	$\frac{6}{5}$
2	$\frac{27}{23}$	$\frac{65}{23}$

$\overset{4}{\alpha}_{3jk}$

j \ k	1	2
1	$\frac{14}{5}$	$\frac{27}{23}$
2	$\frac{6}{5}$	$\frac{65}{23}$

注意，碎权更新的结果受处理样本的顺序的影响，不同顺序得出的结果可能不同. 为了体会这一点，读者可以在例 7.9 中，调换处理 $\mathbf{D}_3$ 和 $\mathbf{D}_4$ 的次序，然后看看最后结果会有什么不同.

7.9 文献介绍

最大似然估计和贝叶斯估计是统计学中两个基本的参数估计方法. 本章针对贝叶斯网对它们进行了介绍，读者若想对它们做一般的了解，可参阅文献(Duda et al.，2001)的第 3 章. 定理 7.1 和定理 7.2 所述的结果对一般的正则指数型分布族也成立，见文献（Kass and Vos.，1997）中定理 2.5.11.

随机缺失假设是由 Rubin（1976）提出的. EM 算法的基本思想早在 20 世纪 50 年代中期已经在一些研究中出现，Dempster 等（1977）对这些思想进行了整理，得到了一个应用广泛的一般算法，其后有大量相关工作，有兴趣的读者可以参阅专著（McLachlan and Krishnan，1997）. 最早把 EM 算法用于贝叶斯网参数估计的是 Lauritzen（1995）. 计算最大似然估计就是要优化对数似然函数 $l(\boldsymbol{\theta} \mid \mathscr{D})$，Russell 等（1995）提出了一个优化 $l(\boldsymbol{\theta} \mid \mathscr{D})$ 的梯度下降（gradient descent）方法. 有趣的是，这个算法与 EM 算法基本一样，不同的只是它每一步需要决定步长. Cowell 等在文献（Cowell et al.，1999）的第 9 章对怎样利用贝叶斯方法对贝叶斯网的参数进行估计也有详细的论述. 本章介绍的碎权更新源自文献（Titterington，1976）. Zhang（1996）指出，利用定理 4.4 所述的无关性可以加快 EM 算法的收敛，改进贝叶斯估计，Bauer 等（1997）讨论了加快 EM 算法收敛的另外一些方法.

第8章 结构学习

参数学习假设已知变量间的定性关系，通过数据分析揭示变量间的定量关系；而结构学习则是要同时揭示变量间的定性和定量关系. 结构学习一般分两步讨论，即模型选择（model selection）和模型优化（model optimization）. 模型选择要回答的问题是用什么样的准则来评判不同模型结构之优劣，而模型优化则是要把最优的模型结构找出来. 本章的前四节讨论模型选择，其余部分讨论模型优化.

8.1 似然函数与模型选择

设 X_1，X_2，…，X_n 为一组随机变量，$\mathscr{D}=(\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m)$是关于这些变量的一组数据. 本章所关心的问题是如何找出一个相对于 $\mathscr{D}$ 在某种意义下最优的贝叶斯网.

设 $\mathscr{G}$ 是一个以 X_1，X_2，…，X_n 为节点的贝叶斯网结构. $\mathscr{G}$ 相对于数据 $\mathscr{D}$ 的优劣可以用一个**评分函数**(scoring function) 来度量. 本节介绍一个基于似然函数的评分准则，称为**最优参数对数似然函数**(parameter maximized loglikelihood function)，简称优参对数似然函数；第 8.2 节在贝叶斯框架下，推出另一个称为 **CH 评分**的准则；第 8.3 节在大样本的前提下得到第三个准则，即 **BIC 评分**；第 8.4 节介绍其它评分准则.

结构 $\mathscr{G}$ 与相应的参数集合 $\boldsymbol{\theta}_{\mathscr{G}}$ 组成贝叶斯网($\mathscr{G}$，$\boldsymbol{\theta}_{\mathscr{G}}$). 这里我们给 $\boldsymbol{\theta}_{\mathscr{G}}$ 加下标是因为不同的网络结构对应不同的参数集合. 例如，在例 7.4 所讨论的结构中，X_3 的父节点包括 X_1 和 X_2，关于 X_3 的条件概率分布的参数有 θ_{311}、θ_{312}、θ_{321}、θ_{322}、θ_{331}、θ_{332}、θ_{341}和 θ_{342}，其中参数 θ_{311}的含义是 $P(X_3=1 \mid X_1=1, X_2=1)$. 但是，如果去掉边 $X_2 \rightarrow X_3$，那么 X_3 就只有 一个父节点 X_1，关于 X_3 的条件概率分布的参数变成 θ_{311}，θ_{312}，θ_{321}和 θ_{322}，其中 θ_{311}的含义是 $P(X_3=1 \mid X_1=1)$.

在贝叶斯网($\mathscr{G}$，$\boldsymbol{\theta}_{\mathscr{G}}$) 中，可以计算每一个样本 $\mathbf{D}_l$ 的概率 $P(\mathbf{D}_l \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})$，从而在 i. i. d. 的假设下，可以计算 $\log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})$. 在参数学习中，网络结构 $\mathscr{G}$ 已知，这就是参数向量 $\boldsymbol{\theta}_{\mathscr{G}}$ 的对数似然函数. 根据最大似然估计原则，相对于数据 $\mathscr{D}$ 最优的参数值 $\boldsymbol{\theta}_{\mathscr{G}}^*$ 应该使 $\log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})$ 达到最大，即

$$\log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}}^*) = \sup_{\boldsymbol{\theta}_{\mathscr{G}}} \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}}).$$

在结构学习中，网络结构 $\mathscr{G}$ 和网络参数 $\boldsymbol{\theta}_{\mathscr{G}}$ 都是需要确定的对象. 于是可以将

$\log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})$ 视为二元组 $(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})$ 的**对数似然函数**，记为

$$l(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D}) = \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}}). \tag{8.1}$$

将最大似然估计原则加以推广，得到如下原则：相对于数据 $\mathscr{D}$ 最优的贝叶斯网 $(\mathscr{G}^*, \boldsymbol{\theta}_{\mathscr{G}^*}^*)$ 应该使对数似然函数达到最大，即

$$l(\mathscr{G}^*, \boldsymbol{\theta}_{\mathscr{G}^*}^* \mid \mathscr{D}) = \max_{\mathscr{G}} \sup_{\boldsymbol{\theta}_{\mathscr{G}}} l(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D}). \tag{8.2}$$

在概念上，寻找最优贝叶斯网的过程可以分成两步：第一步寻找最优结构 $\mathscr{G}^*$，第二步优化参数 $\boldsymbol{\theta}_{\mathscr{G}^*}$. 为刻画第一步所使用的准则，对任一网络结构 $\mathscr{G}$，定义

$$l^*(\mathscr{G} \mid \mathscr{D}) = \sup_{\boldsymbol{\theta}_{\mathscr{G}}} l(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D}). \tag{8.3}$$

作为网络结构 $\mathscr{G}$ 的函数，$l^*(\mathscr{G} \mid \mathscr{D})$ 称为优参对数似然函数. 根据式 (8.2)，最优结构 $\mathscr{G}^*$ 应该使优参对数似然函数达到最大，即

$$l^*(\mathscr{G}^* \mid \mathscr{D}) = \max_{\mathscr{G}} l^*(\mathscr{G} \mid \mathscr{D}). \tag{8.4}$$

这就是我们介绍的第一个模型选择准则，称为**最大优参似然准则**.

下面探讨最大优参似然准则的性质. 首先证明一个关于优参对数似然函数的引理.

引理 8.1 设数据 $\mathscr{D}$ 是完整的，$\hat{P}$ 是 $\mathscr{D}$ 的经验分布，则有

$$l^*(\mathscr{G} \mid \mathscr{D}) = -m \sum_{i=1}^{n} H_{\hat{P}}(X_i \mid \pi_{\mathscr{G}}(X_i)), \tag{8.5}$$

其中 $\pi_{\mathscr{G}}(X_i)$ 是 X_i 在结构 $\mathscr{G}$ 中的父节点的集合，$H_{\hat{P}}(\cdot \mid \cdot)$ 是基于分布 $\hat{P}$、给定 $\pi_{\mathscr{G}}(X_i)$ 时 X_i 的条件熵（见第 1.5.3 节，式 (1.24)）.①

证明 根据式 (7.27)，有

$$l(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D}) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \log \theta_{ijk},$$

其中 q_i，m_{ijk} 和 θ_{ijk} 与 $\mathscr{G}$ 有关. 给定 $\mathscr{G}$，$\boldsymbol{\theta}_{\mathscr{G}}$ 的最大似然估计 $\boldsymbol{\theta}_{\mathscr{G}}^* = \{\theta_{ijk}^*\}$ 由式 (7.28) 给出. 所以，

$$l^*(\mathscr{G} \mid \mathscr{D}) = l(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}}^* \mid \mathscr{D}) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \log \frac{m_{ijk}}{\sum_{k=1}^{r_i} m_{ijk}}.$$

由引理 7.1 可知，

$$m_{ijk} = m\hat{P}(X_i = k, \pi_{\mathscr{G}}(X_i) = j)$$

所以

① 这里假设 $\hat{P}(X_i \mid \pi_{\mathscr{G}}(X_i))$ 是有定义的，即 $\hat{P}(\pi_{\mathscr{G}}(X_i))$ 大于零.

$$l^*(\mathscr{G} \mid \mathscr{D}) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} m\hat{P}(X_i = k, \pi_{\mathscr{G}}(X_i) = j)\log \frac{m\hat{P}(X_i = k, \pi_{\mathscr{G}}(X_i) = j)}{\sum_{k=1}^{r_i} m\hat{P}(X_i = k, \pi_{\mathscr{G}}(X_i) = j)}$$

$$= -m\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} \hat{P}(X_i = k, \pi_{\mathscr{G}}(X_i) = j)\log \frac{1}{\hat{P}(X_i = k \mid \pi_{\mathscr{G}}(X_i) = j)}$$

$$= -m\sum_{i=1}^{n} H_{\hat{P}}(X_i \mid \pi_{\mathscr{G}}(X_i)).$$

引理得证. □

在一个贝叶斯网结构中，如果任何两个节点之间都有一条边，则称该网络是**完全的**（complete）.

定理 8.1 使优参似然函数达到最大的网络结构 $\mathscr{G}^*$ 是完全的.

证明 用反证法. 假若 $\mathscr{G}^*$ 不是完全的，那么一定有两个节点 X_i 和 X_j 之间没有边. 由于 $\mathscr{G}^*$ 是有向无圈图，一定可以在 X_i 和 X_j 之间加一条有向边，使得结果仍是无圈图. 不失一般性，设可以加的边是 $X_j \to X_i$，记所得网络结构为 $\mathscr{G}'$. 根据引理 8.1，得

$$l^*(\mathscr{G}' \mid \mathscr{D}) - l^*(\mathscr{G}^* \mid \mathscr{D}) = -mH_{\hat{P}}(X_i \mid \pi_{\mathscr{G}'}(X_i)) + mH_{\hat{P}}(X_i \mid \pi_{\mathscr{G}^*}(X_i)).$$

根据 $\mathscr{G}'$ 的构造，有 $\pi_{\mathscr{G}}(X_i) = \pi_{\mathscr{G}^*}(X_i) \cup \{X_j\}$. 由定理 1.5，

$$H_{\hat{P}}(X_i \mid \pi_{\mathscr{G}^*}(X_i), X_j) \leqslant H_{\hat{P}}(X_i \mid \pi_{\mathscr{G}^*}(X_i)), \tag{8.6}$$

而且上式中的等号成立当且仅当

$$\hat{P}(X_i \mid \pi_{\mathscr{G}^*}(X_i), X_j) = \hat{P}(X_i \mid \pi_{\mathscr{G}^*}(X_i)). \tag{8.7}$$

由于 $\hat{P}$ 是经验分布，因而受随机性的影响，式（8.7）不成立①. 于是，式（8.6）中的等号也不成立. 进而有

$$l^*(\mathscr{G}' \mid \mathscr{D}) > l^*(\mathscr{G}^* \mid \mathscr{D}).$$

这与 $\mathscr{G}^*$ 的定义矛盾. 定理得证. □

定理 8.1 表明，优参似然函数不适宜作为模型选择的准则，因为无论原生模型如何，使用它都会得到一个完全的有向无圈图. 如果原生模型的结构稀疏（sparse），那么变量之间有许多条件独立关系，使用优参似然函数完全不能够揭示这些关系. 另外，完全模型的复杂度与所有变量的联合分布的复杂度一样高，所以难以处理. 而且，由于所得模型复杂度高、参数众多，会导致**过度拟合**（overfitting）.

---

① 这一步证明不十分严格，严格地讲应该是“式（8.7）成立的概率很小”. 因此，定理 8.1 的陈述也不十分严格. 下面读者会看到，证明该定理的目的是要说明优参似然函数不适宜作为模型选择的准则，而它的陈述和证明的不严谨并不妨碍这个目的. 当然定理 8.1 及其证明可以用严谨的数学语言重写，但相关的技术处理会影响可读性，不利于读者了解其中的要点.

8.2 贝叶斯模型选择

在贝叶斯模型选择框架中，我们视模型结构 $\mathscr{G}$ 和模型参数 $\boldsymbol{\theta}_{\mathscr{G}}$ 为随机变量. 变量 $\mathscr{G}$ 的可能取值包括所有以 X_1，X_2，…，X_n 为节点的有向无圈图．给定 $\mathscr{G}$，变量 $\boldsymbol{\theta}_{\mathscr{G}}$ 的可能取值是所有与 $\mathscr{G}$ 对应的参数值. 我们把关于结构 $\mathscr{G}$ 的先验知识概括为一个概率分布 $P(\mathscr{G})$，称之为**结构先验分布**（structure prior）. 对于一个给定的结构 $\mathscr{G}$，我们把关于参数 $\boldsymbol{\theta}_{\mathscr{G}}$ 的先验知识概括为另一个概率分布 $p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G})$，称之为**参数先验分布**（parameter prior）. 这样，就有了一个关于二元组（$\mathscr{G}$，$\boldsymbol{\theta}_{\mathscr{G}}$）的先验分布

$$p(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}}) = P(\mathscr{G})p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G}). \tag{8.8}$$

在观测到数据 $\mathscr{D}=(\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_m)$ 后，需要计算的是后验概率分布 $p(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D})$. 根据式（1.17），有

$$p(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D}) \propto P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})p(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}}). \tag{8.9}$$

$p(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D})$ 就是关于（$\mathscr{G}$，$\boldsymbol{\theta}_{\mathscr{G}}$）的贝叶斯估计. 基于这个估计，可以对下一个样本 $\mathbf{D}_{m+1}$ 进行预测，即计算其概率分布 $P(\mathbf{D}_{m+1} \mid \mathscr{D})$：

$$\begin{aligned} P(\mathbf{D}_{m+1} \mid \mathscr{D}) &= \sum_{\mathscr{G}} \int P(\mathbf{D}_{m+1} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})p(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D})\mathrm{d}\boldsymbol{\theta}_{\mathscr{G}} \\ &= \sum_{\mathscr{G}} \int P(\mathbf{D}_{m+1} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})P(\mathscr{G} \mid \mathscr{D})p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G}, \mathscr{D})\mathrm{d}\boldsymbol{\theta}_{\mathscr{G}} \\ &= \sum_{\mathscr{G}} P(\mathscr{G} \mid \mathscr{D}) \int P(\mathbf{D}_{m+1} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G}, \mathscr{D})\ \mathrm{d}\boldsymbol{\theta}_{\mathscr{G}}. \end{aligned} \tag{8.10}$$

其中第二个等式成立是因为 $p(\mathscr{G},\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D})=P(\mathscr{G} \mid \mathscr{D})p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G}, \mathscr{D})$，而第三个等式成立是因为 $p(\mathscr{G} \mid \mathscr{D})$ 与 $\boldsymbol{\theta}_{\mathscr{G}}$ 无关. 最后得到的式(8.10)称为 $P(\mathbf{D}_{m+1} \mid \mathscr{D})$ 的完全贝叶斯估计.

为了解读式(8.10)，先来看其中的一个部分 $\int P(\mathbf{D}_{m+1} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G}, \mathscr{D})\mathrm{d}\boldsymbol{\theta}_{\mathscr{G}}$，它是在给定结构 $\mathscr{G}$ 的情况下对 $\mathbf{D}_{m+1}$ 进行完全贝叶斯估计[参见式(7.38)]. 其结果可以表示成一个贝叶斯网，该网络的结构是 $\mathscr{G}$，其参数由式（7.42）给出. 因此，式（8.10）可做如下解读：由于不知道贝叶斯网的网络结构，逐个考虑每一个可能的结构；对每一个可能的结构 $\mathscr{G}$，用贝叶斯方法进行参数估计，得到一个贝叶斯网；最后将获得的所有贝叶斯网的联合概率加权平均，一个贝叶斯网的权重就是其结构 $\mathscr{G}$ 的后验概率 $P(\mathscr{G} \mid \mathscr{D})$；加权平均的结果就是 $\mathbf{D}_{m+1}$ 的分布 $P(\mathbf{D}_{m+1} \mid \mathscr{D})$. 这个计算 $P(\mathbf{D}_{m+1} \mid \mathscr{D})$ 的过程称为**模型平均**（model averaging）.

$\mathbf{D}_{m+1}$ 由随机变量 X_1，X_2，…，X_n 组成. 若已知其中一些变量的取值 $\mathbf{E}=\mathbf{e}$，怎样计算另外一些变量的后验概率 $P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$ 呢？用 $\mathscr{N}_{\mathscr{G}}$ 记上段所述过程中以

$\mathscr{G}$ 为结构的贝叶斯网. 根据式 (8.10)，需要在每个网络 $\mathscr{N}_{\mathscr{G}}$ 中计算 $P_{\mathscr{N}_{\mathscr{G}}}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e})$，然后将结果加权平均，即

$$P(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}) = \sum_{\mathscr{G}} P(\mathscr{G} \mid \mathscr{D}) P_{\mathscr{N}_{\mathscr{G}}}(\mathbf{Q} \mid \mathbf{E}=\mathbf{e}). \tag{8.11}$$

给定一组变量，以它们为节点的有向无圈图有很多，难以一一列举（参见 8.5 节）. 所以，在实际中模型平均只会考虑为数不多的几个模型. 最常见的是考虑对应后验概率 $P(\mathscr{G} \mid \mathscr{D})$ 最大的那个模型，即

$$\mathscr{G}^* = \arg\max_{\mathscr{G}} P(\mathscr{G} \mid \mathscr{D}). \tag{8.12}$$

这样，利用贝叶斯模型选择进行结构学习的实质过程是：

(1) 找出后验概率最大的结构 $\mathscr{G}^*$；

(2) 相对于 $\mathscr{G}^*$ 进行贝叶斯参数估计.

由于 $P(\mathscr{G} \mid \mathscr{D}) = \dfrac{P(\mathscr{D} \mid \mathscr{G})\ P(\mathscr{G})}{P(\mathscr{D})}$，而 $P(\mathscr{D})$ 不依赖于 $\mathscr{G}$，所以选择后验概率最大的结构就是选择使如下函数达到最大的结构：

$$\log P(\mathscr{G}, \mathscr{D}) = \log P(\mathscr{D} \mid \mathscr{G}) + \log P(\mathscr{G}). \tag{8.13}$$

$\log P(\mathscr{G}, \mathscr{D})$ 称为结构 $\mathscr{G}$ 的**贝叶斯评分** (Bayesian score). 式 (8.13) 右边第二项中的 $P(\mathscr{G})$ 是结构先验分布，一般假设它是均匀分布. 展开右边第一项中的 $P(\mathscr{D} \mid \mathscr{G})$，有

$$P(\mathscr{D} \mid \mathscr{G}) = \int P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}}) p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G}) \mathrm{d}\boldsymbol{\theta}_{\mathscr{G}}. \tag{8.14}$$

这里 $P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})$ 是二元组 $(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}})$ 的似然函数，记为 $L(\mathscr{G}, \boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{D})$. 因此，$P(\mathscr{D} \mid \mathscr{G})$称为**边缘似然函数** (marginal likelihood)，记为 $L(\mathscr{G} \mid \mathscr{D})$. 注意，它与式 (8.3) 所定义的优参似然函数很不一样：前者对 $\boldsymbol{\theta}_{\mathscr{G}}$ 积分，后者取 $\boldsymbol{\theta}_{\mathscr{G}}$ 的最优值 $\boldsymbol{\theta}_{\mathscr{G}}{}^*$.

定理 8.2 设 $\mathscr{D}$ 是一组关于变量 X_1，X_2，…，X_n 的完整 i. i. d. 数据，$\mathscr{G}$ 是一个以 X_1，X_2，…，X_n 为节点的贝叶斯网结构. 如果参数先验分布 $p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G})$是如下的乘积狄利克雷分布：

$$p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G}) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1}, \tag{8.15}$$

那么，

$$L(\mathscr{G} \mid \mathscr{D}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*}+m_{ij*})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk}+m_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{8.16}$$

其中 m_{ijk} 是 $\mathscr{D}$ 中满足 $X_i = k$，$\pi(X_i) = j$ 的样本个数，$m_{ij*} = \sum_{k=1}^{r_i} m_{ijk}$，$\alpha_{ij*} = \sum_{k=1}^{r_i} \alpha_{ijk}$.

证明 为方便起见，约定在 $\mathscr{D}=\phi$ 时，$P(\mathscr{D}|\mathscr{G})=P(\phi|\mathscr{G})=1$. 在这一约定下，下式即使在 $\mathscr{D}=\phi$ 时也成立：

$$P(\mathscr{D},\mathbf{D}_{m+1}\mid\mathscr{G})=P(\mathscr{D}\mid\mathscr{G})P(\mathbf{D}_{m+1}\mid\mathscr{G},\mathscr{D}). \tag{8.17}$$

我们通过对样本个数进行归纳来证明式（8.16）. 当 $\mathscr{D}=\phi$ 时，式（8.16）等号左边按约定是1. 而由于 $m_{ijk}=0$，等号右边也是1. 所以，此时式（8.16）成立. 设式（8.16）在 m 个样本即 $\mathscr{D}=(\mathbf{D}_1,\mathbf{D}_2,\cdots,\mathbf{D}_m)$ 的情况下成立. 下面证明它在 $m+1$ 个样本即 $\mathscr{D}\cup\mathbf{D}_{m+1}=(\mathbf{D}_1,\mathbf{D}_2,\cdots,\mathbf{D}_m,\mathbf{D}_{m+1})$时也成立. 设 m_{ijk}是 m 个样本时，满足 $X_i=k$，$\pi(X_i)=j$ 的样本的个数；m'_{ijk}是 $m+1$ 个样本时，满足 $X_i=k$，$\pi(X_i)=j$ 的样本的个数. 根据式（7.26），有

$$m_{ijk}=\sum_{l=1}^{m}\chi(i,j,k:\mathbf{D}_l),\ m'_{ijk}=\sum_{l=1}^{m+1}\chi(i,j,k:\mathbf{D}_l).$$

于是

$$m'_{ijk}=m_{ijk}+\chi(i,j,k:\mathbf{D}_{m+1}),\ m'_{ij*}=m_{ij*}+\chi(i,j,*:\mathbf{D}_{m+1}). \tag{8.18}$$

其中 $\chi(i,j,*:\mathbf{D}_{m+1})=\sum_{k=1}^{r_i}\chi(i,j,k:\mathbf{D}_{m+1})$.

考虑式（8.17）. 根据归纳假设，有

$$P(\mathscr{D}\mid\mathscr{G})=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*}+m_{ij*})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+m_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{8.19}$$

而 $P(\mathbf{D}_{m+1}\mid\mathscr{G},\mathscr{D})$ 是在给定网络结构 $\mathscr{G}$ 的情况下，基于 $\mathscr{D}$ 对 $\mathbf{D}_{m+1}$的分布进行贝叶斯估计的结果. 这个结果由式（7.43）给出. 不难看出，它可以改写成如下形式：

$$P(\mathbf{D}_{m+1}\mid\mathscr{G},\mathscr{D})=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{1}{(m_{ij*}+\alpha_{ij*})^{\chi(i,j,*:\mathbf{D}_{m+1})}}\prod_{k=1}^{r_i}(m_{ijk}+\alpha_{ijk})^{\chi(i,j,k:\mathbf{D}_{m+1})}. \tag{8.20}$$

将式（8.19）和式（8.20）带入式（8.17），并稍做简化，得

$$P(\mathscr{D},\mathbf{D}_{m+1}\mid\mathscr{G})=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*}+m_{ij*})\,(m_{ij*}+\alpha_{ij*})^{\chi(i,j,*:\mathbf{D}_{m+1})}}\times\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+m_{ijk})\,(m_{ijk}+\alpha_{ijk})^{\chi(i,j,k:\mathbf{D}_{m+1}}}{\Gamma(\alpha_{ijk})}.$$

由于 $\chi(i,j,*:\mathbf{D}_{m+1})$ 非0即1，而Γ-函数满足 $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$，有

$$P(\mathscr{D},\mathbf{D}_{m+1}\mid\mathscr{G})=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*}+m_{ij*}+\chi(i,j,*:\mathbf{D}_{m+1}))}\times\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+m_{ijk}+\chi(i,j,k:\mathbf{D}_{m+1}))}{\Gamma(\alpha_{ijk})}.$$

利用式（8.18），得

$$P(\mathscr{D},\mathbf{D}_{m+1} \mid \mathscr{G}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*}+m'_{ij*})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+m'_{ijk})}{\Gamma(\alpha_{ijk})}.$$

这就证明了式（8.16）在 $m+1$ 个样本时也成立．所以，定理得证． □

对式（8.16）两边取对数，得

$$l(\mathscr{G} \mid \mathscr{D}) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log\frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*}+m_{ij*})}+\sum_{k=1}^{r_i}\log\frac{\Gamma(\alpha_{ijk}+m_{ijk})}{\Gamma(\alpha_{ijk})}\right]. \quad (8.21)$$

式（8.21）右边所给出的量称为结构 $\mathscr{G}$ 的 **Cooper-Herskovits 评分**，简称 **CH 评分**．如果假设结构先验分布是均匀分布，那么用贝叶斯评分选择模型就等于是用 CH 评分来选择模型．

在使用 CH 评分之前，首先需要选定参数先验分布 $p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G})$ 中的超参数 α_{ijk}．通常这并非易事，因为理论上我们需要对每一个可能的结构都提供参数先验分布，而结构数目众多，无法一一罗列．在实际中，人们往往规定一个等价样本量 α 和一个先验贝叶斯网 $\mathscr{N}_0$，并且利用下式得到 $p(\boldsymbol{\theta}_{\mathscr{G}} \mid \mathscr{G})$ 的超参数 α_{ijk}：

$$\alpha_{ijk} = \alpha P_{\mathscr{N}_0}(X_i = k \mid \pi_{\mathscr{G}}(X_i) = j). \quad (8.22)$$

8.3 大样本模型选择

下面介绍另一个模型评分函数，即**贝叶斯信息准则**（Bayesian information criterion），简称 **BIC 评分**．BIC 评分是在大样本前提下对边缘似然函数的一种近似．它有明确直观的意义，而且使用方便，是实际中最常用的评分函数．

本节主要利用**拉普拉斯近似**（Laplace approximation）方法，对 $P(\mathscr{D} \mid \mathscr{G})$ 进行大样本近似，从而推导出 BIC 评分函数．基本的想法是在最大似然估计附近把对数似然函数展开，然后将计算转化为一个多元正态分布函数在极值点的邻域的积分．我们从式（8.14）开始．为简化记号，把 $\boldsymbol{\theta}_{\mathscr{G}}$ 改记为 $\boldsymbol{\theta}$．用 $\boldsymbol{\theta}^* = \{\theta^*_{ijk} \mid i=1,2,\cdots,n; j=1,2,\cdots,q_i; k=1,2,\cdots,r_i\}$ 记 $\boldsymbol{\theta}$ 的最大似然估计．注意，q_i 依赖于模型结构 $\mathscr{G}$．关于 $\boldsymbol{\theta}^*$，假设它是 $\mathscr{G}$ 的参数空间的一个**内点**（interior point），即对所有 i,j,k，

$$\theta^*_{ijk} > 0. \quad (8.23)$$

这样，$\boldsymbol{\theta}^*$ 作为 $P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta})$ 的最大值点是被式（7.28）唯一确定的．关于先验参数分布 $p(\boldsymbol{\theta} \mid \mathscr{G})$，假设它在 $\boldsymbol{\theta}^*$ 周围是光滑的并且不为零．最后，还假设数据是完整的．

为了考察 $P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta})$ 的性质，简记 $\log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta})$ 为 $l(\boldsymbol{\theta})$．根据式（7.27）和式（7.28），有

$$l(\boldsymbol{\theta})=\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} m_{ijk}\log\theta_{ijk} \tag{8.24}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{q_i} m_{ij*}\left[\sum_{k=1}^{r_i}\theta_{ijk}^{*}\log\frac{\theta_{ijk}}{\theta_{ijk}^{*}}+\sum_{k=1}^{r_i}\theta_{ijk}^{*}\log\theta_{ijk}^{*}\right]. \tag{8.25}$$

于是，

$$P(\mathscr{D}\mid\mathscr{G},\boldsymbol{\theta})=\prod_{i=1}^{n}\prod_{j=1}^{q_i}\left(\exp\left\{\sum_{k=1}^{r_i}\theta_{ijk}^{*}\log\frac{\theta_{ijk}}{\theta_{ijk}^{*}}+\sum_{k=1}^{r_i}\theta_{ijk}^{*}\log\theta_{ijk}^{*}\right\}\right)^{m_{ij*}}. \tag{8.26}$$

根据假设，我们已经知道，作为一个 $\boldsymbol{\theta}$ 的函数，$P(\mathscr{D}\mid\mathscr{G},\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}^*$ 处达到唯一的最大值. 另一方面，在样本量 m 很大时，m_{ij*} 也很大. 根据式 (8.26)，当 $\boldsymbol{\theta}$ 与 $\boldsymbol{\theta}^*$ 之间距离的增加时，$P(\mathscr{D}\mid\mathscr{G},\boldsymbol{\theta})$的取值迅速降低. 又因为 $p(\boldsymbol{\theta}\mid\mathscr{G})$在 $\boldsymbol{\theta}^*$ 周围光滑且不为零，所以，$P(\mathscr{D}\mid\mathscr{G})=\int P(\mathscr{D}\mid\mathscr{G},\boldsymbol{\theta})p(\boldsymbol{\theta}\mid\mathscr{G})\,\mathrm{d}\boldsymbol{\theta}$ 可以用 $P(\mathscr{D}\mid\mathscr{G},\boldsymbol{\theta})p(\boldsymbol{\theta}\mid\mathscr{G})$在 $\boldsymbol{\theta}^*$的一个小邻域 $\mathrm{nb}(\boldsymbol{\theta}^*)$中的积分来近似，即

$$P(\mathscr{D}\mid\mathscr{G})\approx\int_{\mathrm{nb}(\boldsymbol{\theta}^*)}P(\mathscr{D}\mid\mathscr{G},\boldsymbol{\theta})p(\boldsymbol{\theta}\mid\mathscr{G})\mathrm{d}\boldsymbol{\theta}, \tag{8.27}$$

在 $\boldsymbol{\theta}^*$ 周围将 $l(\boldsymbol{\theta})$进行泰勒展开，得在邻域 $\mathrm{nb}(\boldsymbol{\theta}^*)$内，有

$$l(\boldsymbol{\theta})\approx l(\boldsymbol{\theta}^*)+\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\mathrm{T}}l''(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*), \tag{8.28}$$

其中 $l''(\boldsymbol{\theta}^*)$是 $l(\boldsymbol{\theta})$的**黑塞矩阵** (Hessian matrix) 在 $\boldsymbol{\theta}^*$ 的值：

$$l''(\boldsymbol{\theta}^*)=\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_{ijk}\partial\,\theta_{i'j'k'}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{8.29}$$

以下用 A 记$-l''(\boldsymbol{\theta}^*)$. 因为式 (8.24)，$l(\boldsymbol{\theta})$是一个凹函数，从而 A 是正定的. 根据 $p(\boldsymbol{\theta}\mid\mathscr{G})$在 $\boldsymbol{\theta}^*$ 附近的光滑性，在小邻域 $\mathrm{nb}(\boldsymbol{\theta}^*)$内，有

$$p(\boldsymbol{\theta}\mid\mathscr{G})\approx p(\boldsymbol{\theta}^*\mid\mathscr{G}). \tag{8.30}$$

将式 (8.28) 和式 (8.30) 代入式 (8.27) 得

$$\begin{aligned}P(\mathscr{D}\mid\mathscr{G})&\approx\int\exp\{l(\boldsymbol{\theta})\}\,p(\boldsymbol{\theta}\mid\mathscr{G})\mathrm{d}\boldsymbol{\theta}\\&\approx\int_{\mathrm{nb}(\boldsymbol{\theta}^*)}\exp\left\{l(\boldsymbol{\theta}^*)-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\mathrm{T}}A(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right\}p(\boldsymbol{\theta}^*\mid\mathscr{G})\mathrm{d}\boldsymbol{\theta}\\&=\exp\{l(\boldsymbol{\theta}^*)\}p(\boldsymbol{\theta}^*\mid\mathscr{G})\\&\quad\times\int_{\mathrm{nb}(\boldsymbol{\theta}^*)}\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\mathrm{T}}A(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right\}\mathrm{d}\boldsymbol{\theta}.\end{aligned} \tag{8.31}$$

注意，参数向量 $\boldsymbol{\theta}$ 由 $d=\prod_{i=1}^{n}q_i(r_i-1)$ 个独立参数组成. 由于以 A 为协方差矩阵、以 $\boldsymbol{\theta}^*$ 为均值的正态分布的密度函数是

$$\frac{1}{\sqrt{(2\pi)^d \mid A \mid^{-1}}}\exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\mathrm{T}}A(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right\},$$

而且 $\exp\{l(\boldsymbol{\theta}^*)\}=P(\mathscr{D} \mid \mathscr{G},\boldsymbol{\theta}^*)$，所以，

$$P(\mathscr{D} \mid \mathscr{G}) \approx P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*)p(\boldsymbol{\theta}^* \mid \mathscr{G})\sqrt{(2\pi)^d \mid A \mid^{-1}}. \tag{8.32}$$

于是有

$$\log P(\mathscr{D} \mid \mathscr{G}) \approx \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*) - \frac{1}{2}\log \mid A \mid + \log p(\boldsymbol{\theta}^* \mid \mathscr{G}) + \frac{d}{2}\log(2\pi). \tag{8.33}$$

式（8.33）称为 $\log P(\mathscr{D} \mid \mathscr{G})$的拉普拉斯近似.

拉普拉斯近似的后两项不依赖于样本量 m，一般将其略去. 而 $\log \mid A \mid$ 可以用 $d \log m$ 来近似（Raftery，1995），于是得

$$\log P(\mathscr{D} \mid \mathscr{G}) \approx \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*) - \frac{d}{2}\log m. \tag{8.34}$$

这就是模型结构 $\mathscr{G}$ 的 BIC 评分，记为 $\mathrm{BIC}(\mathscr{G} \mid \mathscr{D})$.

BIC 评分的第一项是模型 $\mathscr{G}$ 的优参对数似然度，它度量的是结构 $\mathscr{G}$ 与数据 $\mathscr{D}$ 的拟合程度. 第二项是一个关于模型复杂度的罚项（penalty）. 在第 8.1 节中我们已经看到，若仅仅依据优参似然度来选择模型，会选到最复杂的完全贝叶斯网络，从而导致过度拟合. 由于附加了一个模型复杂度的罚项，BIC 有效地避免了过度拟合. 直观上，基于 BIC 评分选择模型就是要选择既与数据拟合、又比较简单的模型.

8.4 其它模型选择标准

除前面讨论的三个评分函数以外，还有另外几个模型选择准则也被用于贝叶斯网的结构学习，这包括 MDL 评分、AIC 评分、验证数据似然度以及交叉验证. 本节先对这几个准则进行简单介绍，然后还介绍一个关于模型选择准则的重要性质，即一致性.

MDL 评分是**最短描述长度**（minimum description length）(Rissanen，1978) 的简称. 这个准则的基本思想如下：数据分析的目的是要找出蕴含在数据中的规律，然后可以利用它们对数据进行压缩，从而降低数据的编码（描述）长度. 所以，用贝叶斯网分析数据是否成功可以用数据和模型的编码总长度来度量.

MDL 评分是这个思想的具体化. 在具体化的过程中，需要做一些假设和近似. 不同的假设和近似导致不同的 MDL 评分函数. 在大样本的情况下，用这些

MDL 评分函数进行模型选择的结果基本上相互等价，而且它们也与 BIC 评分等价（Lanterman，2001）.

AIC 评分是 **Akaike 信息准则**（Akaike information criterion）（Akaike，1973）的简称. 它假设数据 $\mathscr{D}$ 是从一个概率分布 $P(\mathbf{X})$ 中进行 i. i. d 抽样而得到的. AIC 评分的出发点是要找一个贝叶斯网 $\mathscr{N}^*$，使得 $P_{\mathscr{N}^*}(\mathbf{X})$ 与 $P(\mathbf{X})$ 之间的 KL 距离最短，即

$$\mathrm{KL}(P,\ P_{\mathscr{N}^*}) \leqslant \mathrm{KL}(P, P_{\mathscr{N}}),\ \forall \mathscr{N}, \tag{8.35}$$

在一定光滑条件下做大样本近似，可得如下结论，即 $\mathscr{N}^*$ 的结构 $\mathscr{G}^*$ 应该满足：

$$\mathrm{AIC}(\mathscr{G}^* \mid \mathscr{D}) \geqslant \mathrm{AIC}(\mathscr{G} \mid \mathscr{D}),\ \forall \mathscr{G}, \tag{8.36}$$

其中，

$$\mathrm{AIC}(\mathscr{G} \mid \mathscr{D}) = \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*) - d. \tag{8.37}$$

这就是 **AIC 评分**.

AIC 评分与 BIC 评分都是优参对数似然度加一个罚项，因此都称为**罚项似然度**（penalized likelihood）（Green，1998）. 另外，MDL 也是罚项似然度.

罚项的作用是防止过度拟合. 下面介绍另外一种防止过度拟合的方法. 它的基本思想是把数据 $\mathscr{D}$ 随机地分成**训练数据**（training data）$\mathscr{D}_t$ 和**验证数据**（validation data）$\mathscr{D}_v$. 对于一个模型结构 $\mathscr{G}$，首先基于训练数据对其参数进行估计，得到一个贝叶斯网（$\mathscr{G}$，$\boldsymbol{\theta}^t$），然后计算验证数据 $\mathscr{D}_v$ 对数似然度：

$$\mathrm{HVL}(\mathscr{G} \mid \mathscr{D}_v, \mathscr{D}_t) = \log P(\mathscr{D}_v \mid \mathscr{G}, \boldsymbol{\theta}^t). \tag{8.38}$$

这称为 $\mathscr{G}$ 的**验证数据似然度**（holdout validation likelihood），简称 **HVL 评分**. 在大样本的情况下，HVL 准则与 AIC 准则等价（Cowell et al.，1999）.

交叉验证（cross validation）是上述思想的进一步发展（Stone，1977）. 它的基本思想是多次计算模型的 HVL 评分，而每次都按不同方式将 $\mathscr{D}$ 划分为 $\mathscr{D}_t$ 和 $\mathscr{D}_v$，然后计算各次所得评分的平均值，并将其作为模型的最后评分. 这个评分称为 **CVL 评分**. 它比 HVL 评分更具鲁棒性. 但其计算复杂度也高出 HVL 评分数倍. 在大样本情况下，CVL 准则与 AIC 准则等价（Cowell et al.，1999）.

下面讨论关于模型选择标准的一个重要性质，即**一致性**（consistency）. 为此需要引入一些概念. 设 $\mathscr{G}$ 是一个贝叶斯网络结构，P 是一个联合概率分布. 如果存在一个参数值 $\boldsymbol{\theta}$，使得贝叶斯网（$\mathscr{G}$，$\boldsymbol{\theta}$）的联合分布正好是 P，那么说 $\mathscr{G}$ **包含**（contains）P. 如果其它任何一个包含 P 的模型的自由参数个数都不少于 $\mathscr{G}$ 的自由参数个数，那么称 $\mathscr{G}$ 为包含 P 的**最简模型**（parsimonious model）. 一个模型结构 $\mathscr{G}$ 包含另一个 模型结构 $\mathscr{G}'$，如果 $\mathscr{G}$ 包含任何一个以 $\mathscr{G}'$ 为结构的贝叶斯网的联合分布. 若 $\mathscr{G}'$ 包含 $\mathscr{G}$ 且 $\mathscr{G}'$ 包含 $\mathscr{G}$，则称 $\mathscr{G}$ 与 $\mathscr{G}'$ **分布等价**（distribu-

tionally equivalent). 若 $\mathscr{G}$ 与 $\mathscr{G}'$ 分布等价，且具有相同数量的自由参数，则称 $\mathscr{G}$ 与 $\mathscr{G}'$ **等价** (equivalent).

设 $\mathscr{D}$ 是从联合分布 P 中抽样而得到的一组数据. 一个模型评分函数 f 称为是**一致的**，如果对任何 P，当样本量趋于无穷时，下列两个条件得到满足：

(1) 如果 $\mathscr{G}$ 包含 P，而 $\mathscr{G}'$ 不包含 P，则

$$f(\mathscr{G} \mid \mathscr{D}) > f(\mathscr{G}' \mid \mathscr{D}).$$

(2) 如果 $\mathscr{G}$ 和 $\mathscr{G}'$ 都包含 P，且 $\mathscr{G}$ 的自由参数少于 $\mathscr{G}'$ 的自由参数，则

$$f(\mathscr{G} \mid \mathscr{D}) > f(\mathscr{G}' \mid \mathscr{D}).$$

为了体会一致性的重要性，考虑一个贝叶斯网 $\mathscr{N}=(\mathscr{G},\boldsymbol{\theta})$ 及其联合分布 $P(\mathbf{X} \mid \mathscr{G},\boldsymbol{\theta})$. 假设 $\mathscr{G}$ 是包含 $P(\mathbf{X} \mid \mathscr{G}, \boldsymbol{\theta})$ 的最简单模型. 如果评分函数 f 是一致的，那么在样本量趋于无穷时，$\mathscr{G}$ 的评分大于其它任何一个不与 $\mathscr{G}$ 等价的模型 $\mathscr{G}'$，即

$$f(\mathscr{G} \mid \mathscr{D}) > f(\mathscr{G}' \mid \mathscr{D}). \tag{8.39}$$

所以，原则上可以利用 f 得到 $\mathscr{G}$ 或者与 $\mathscr{G}$ 等价的一个模型结构 $\mathscr{G}^*$. 再由定理 7.3，原则上还可以得到 $\mathscr{G}^*$ 的参数 $\boldsymbol{\theta}^*$，使得当样本量趋于无穷时，$P(\mathbf{X} \mid \mathscr{G}^*, \boldsymbol{\theta}^*)$ 几乎必然收敛到 $P(\mathbf{X} \mid \mathscr{G}, \boldsymbol{\theta})$. 这就是说，原则上可以获得数据 $\mathscr{D}$ 的原生模型或与其等价的一个模型.

在前面讨论的模型评分函数中，贝叶斯评分、边缘似然度、BIC 评分和 MDL 评分是一致的；当样本量趋于无穷时，AIC 准则退化为最大优参似然准则，它们都会选择完全贝叶斯网模型，所以 AIC 评分不是一致的；又由于大样本情况下 HVL 和 CVL 准则与 AIC 准则等价，从而它们也不是一致的.

8.5 模型优化

在选定模型评分函数后，接下来是要找出评分最高的网络结构，这就是模型优化. 最直接了当的方法是穷举法，即逐一计算每一个结构的评分，然后选出分数最高的结构. 但是，穷举法在实际中不可行，因为网络结构数目太多. 设 $f(n)$ 是所有由 n 个节点组成的有向无圈图的个数. Robinson (1977) 证明了关于 $f(n)$ 的如下递归结果：

$$f(1)=1,$$

$$f(n)=\sum_{i=1}^{n}(-1)^{i+1}\frac{n!}{(n-i)!i!}f(n-i), n>1.$$

当 $n=10$ 时，$f(10) \approx 4.2\times 10^{18}$. 给定一个有向无圈图，可以用一组变量 $X_1, \cdots, X_n$ 将图中的节点命名，从而得到一个贝叶斯网结构. 由于有多种命名方法，

以 $X_1,\cdots,X_n$ 为节点的贝叶斯网结构的数目远大于 $f(n)$. 所以，穷举所有可能的贝叶斯网结构是不可行的.

本节介绍两个启发式搜索算法. 搜索过程就是计算和比较不同模型结构评分的过程. 下面首先指出 CH 评分和 BIC 评分的一个性质，它能大大降低搜索过程中计算模型评分的运算复杂度.

8.5.1 评分函数的分解

考察式 (8.21) 所给出的 CH 评分. 它可以视为 n 项之和，每一项对应一个变量. 用 $\langle X_i, \pi(X_i)\rangle$ 记由变量 X_i、其父节点 $\pi(X_i)$以及它们之间的边所形成的**局部结构**. 式 (8.21) 中与 X_i 所对应的那一项依赖于局部结构$\langle X_i, \pi(X_i)\rangle$，因此记为 $\mathrm{CH}(\langle X_i, \pi(X_i)\rangle \mid \mathscr{D})$，即

$$\mathrm{CH}(\langle X_i, \pi(X_i)\rangle \mid \mathscr{D}) = \sum_{j=1}^{q_i}\left[\log\frac{\Gamma(\alpha_{ij*})}{\Gamma(\alpha_{ij*}+m_{ij*})}+\sum_{k=1}^{r_i}\log\frac{\Gamma(\alpha_{ijk}+m_{ijk})}{\Gamma(\alpha_{ijk})}\right]. \tag{8.40}$$

$\mathrm{CH}(\langle X_i, \pi(X_i)\rangle \mid \mathscr{D})$ 称为 X_i 的**家族 CH 评分**. 利用家族 CH 评分，式 (8.21) 可以改写为

$$\mathrm{CH}(\mathscr{G} \mid \mathscr{D}) = \sum_{i=1}^{n}\mathrm{CH}(\langle X_i,\pi(X_i)\rangle \mid \mathscr{D}), \tag{8.41}$$

式中右边的 $\pi(X_i)$ 依赖于网络结构 $\mathscr{G}$. 在式 (8.41) 的意义下，我们称 CH 评分是**可分解的**，它分解为各变量的家族 CH 评分之和.

BIC 评分也是可分解的. 将式 (7.27) 和式 (7.28) 代入式 (8.34)，得

$$\mathrm{BIC}(\mathscr{G} \mid \mathscr{D}) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} m_{ijk}\log\frac{m_{ijk}}{m_{ij*}} - \sum_{i=1}^{n}\frac{q_i(r_i-1)}{2}\log m. \tag{8.42}$$

定义变量 X_i 的**家族 BIC 评分**为

$$\mathrm{BIC}(\langle X_i, \pi(X_i)\rangle \mid \mathscr{D}) = \sum_{j=1}^{q_i}\sum_{k=1}^{r_i} m_{ijk}\log\frac{m_{ijk}}{m_{ij*}} - \frac{q_i(r_i-1)}{2}\log m, \tag{8.43}$$

则有

$$\mathrm{BIC}(\mathscr{G} \mid \mathscr{D}) = \sum_{i=1}^{n}\mathrm{BIC}(\langle X_i, \pi(X_i)\rangle \mid \mathscr{D}). \tag{8.44}$$

因此，BIC 评分分解为各变量的家族 BIC 评分之和. 另外，MDL，AIC，HVL 及 CVL 评分都是可分解的.

在搜索最优模型的过程中，每一步都会将当前模型结构 $\mathscr{G}$ 略加修改，得到另一个 结构 $\mathscr{G}'$，然后再计算和比较两者的评分. 通常 $\mathscr{G}$ 和 $\mathscr{G}'$ 的差别不大，只有一两个变量的父节点发生了变化. 如果评分函数是可分解的，那么 $\mathscr{G}$ 与 $\mathscr{G}'$ 评分

的差别只是那一两个变量的家族评分的变化，从而容易计算. 作为一个例子，假设 $\mathscr{G}$ 与 $\mathscr{G}'$ 的唯一区别是某变量 X_i 的父节点发生了变化，从 $\pi(X_i)$ 变成了 $\pi'(X_i)$，同时假设使用的是 BIC 评分，则有

$$\begin{aligned}&\mathrm{BIC}(\mathscr{G}' \mid \mathscr{D}) - \mathrm{BIC}(\mathscr{G} \mid \mathscr{D})\\&\qquad = \mathrm{BIC}(\langle X_i, \pi'(X_i) \mid \mathscr{D}\rangle) - \mathrm{BIC}(\langle X_i, \pi(X_i) \mid \mathscr{D}\rangle).\end{aligned} \tag{8.45}$$

换个角度看，若已知 $\mathrm{BIC}(\mathscr{G} \mid \mathscr{D})$，则可用下式计算 $\mathrm{BIC}(\mathscr{G}' \mid \mathscr{D})$：

$$\begin{aligned}\mathrm{BIC}(\mathscr{G}' \mid \mathscr{D}) = &\mathrm{BIC}(\mathscr{G} \mid \mathscr{D}) + \mathrm{BIC}(\langle X_i, \pi'(X_i) \mid \mathscr{D}\rangle)\\&- \mathrm{BIC}(\langle X_i, \pi(X_i) \mid \mathscr{D}\rangle).\end{aligned} \tag{8.46}$$

8.5.2 K2 算法

K2（Cooper and Herskovits，1992）算法是最早的贝叶斯网结构学习算法之一. 设 $\mathscr{D}$ 是关于变量 $X_1, X_2, \cdots, X_n$ 的一组 i. i. d. 完整数据. K2 算法的目的是要寻找 CH 评分高的模型. 出于计算复杂度的考虑，K2 不是要寻找出 CH 评分最高的模型，而是要寻找在一定条件下的最优模型. 所使用的条件有两个，它们使搜索空间大大缩小.

K2 算法用一个变量排序 ρ 和一个正整数 u 来限制搜索空间，它要寻求的是满足如下两个条件的最优模型 $\mathscr{G}$：

（1）$\mathscr{G}$ 中任一变量的父节点个数不超过 u；

（2）ρ 是 $\mathscr{G}$ 的一个拓扑序.

为简化模型评分的计算，K2 假设所有参数先验分布都是均匀分布. 这意味着式（8.40）中的 α_{ijk} 以及 α_{ij*} 都是 1.

K2 算法由图 8.1 给出. 它的出发点是一个包含所有节点、但却没有边的**无边图**（edgeless graph）. 在搜索过程中，K2 按顺序逐个考察 ρ 中的变量，确定其父节点，然后添加相应的边. 对某一变量 X_j，假设 K2 已经找到了它的一些父节点 π_j. 如果 $|\pi_j| < u$，即 X_j 的父节点个数 $|\pi_j|$ 还未达到 u，那么就要继续为它寻找父节点. 具体做法是，首先考虑那些在 ρ 中排在 X_j 之前、但却还不是 X_j 的父节点的变量，从这些变量中选出 X_i，它使得新家族 CH 评分 $V_{\text{new}} = \mathrm{CH}(\langle X_j, \pi_j \cup \{X_i\}\rangle \mid \mathscr{D})$ 达到最大；然后将 V_{new} 与旧家族 CH 评分 $V_{\text{old}} = \mathrm{CH}(\langle X_j, \pi_j\rangle \mid \mathscr{D})$ 比较：如果 $V_{\text{new}} > V_{\text{old}}$，则把 X_i 添加为 X_j 的父节点；否则停止为 X_j 寻找父节点.

K2 是否能够找到高质量的模型结构呢？Cooper 和 Herskovits（1992）为此进行了仿真实验. 实验采用 ALARM 网络，它有 37 个节点和 46 条边，由 Beinlich等（1989）手工构造，其结构如图 8.2 所示，定量信息可以在本书的网页中找到.

K2($\mathbf{X}, \rho, u, \mathscr{D}$)

输入：$\mathbf{X}=\{X_1, X_2, \cdots, X_n\}$——一组变量；

ρ——一个变量顺序(设它与变量下标一致)；

u——变量父节点个数的上界；

$\mathscr{D}$——一组完整数据.

输出：一个贝叶斯网；

1：$\mathscr{G} \leftarrow$ 由节点 $X_1, X_2, \cdots, X_n$ 组成的无边图；
2：**for** j =1 **to** n
3：　$\pi_j \leftarrow \phi$；
4：　$V_{\text{old}} \leftarrow \text{CH}(\langle X_j, \pi_j \rangle | \mathscr{D})$；
5：**while** (**true**)
6：　　$i \leftarrow \arg\max_{1 \leqslant i < j, X_i \notin \pi_j} \text{CH}(\langle X_j, \pi_j \cup \{X_i\} \rangle | \mathscr{D})$
7：　　$V_{\text{new}} \leftarrow \text{CH}(\langle X_j, \pi_j \cup \{X_i\} \rangle | \mathscr{D})$
8：　　**if** ($V_{\text{new}} > V_{\text{old}}$ **and** $|\pi_j| < u$)
9：　　$V_{\text{old}} \leftarrow V_{\text{new}}$；
10：　　$\pi_j \leftarrow \pi_j \cup \{X_i\}$；
11：　　在 $\mathscr{G}$ 中加边 $X_i \rightarrow X_j$；
12：　　**else**
13：　　**break**；
14：　　**end if**
15：**end while**
16：**end for**
17：按式(7.42) 估计 $\mathscr{G}$ 的参数 $\boldsymbol{\theta}$；
18：**return**($\mathscr{G}$, $\boldsymbol{\theta}$)；

图 8.1　学习贝叶斯网结构的 K2 算法（注意 ρ 与变量下标一致的意思是：若 $i<j$，则在 ρ 中 X_i 排列在 X_j 之前. 另外，在第 17 行，估计 $\mathscr{G}$ 的参数 $\boldsymbol{\theta}$ 需要先验参数分布的超参数 α_{ijk}.）

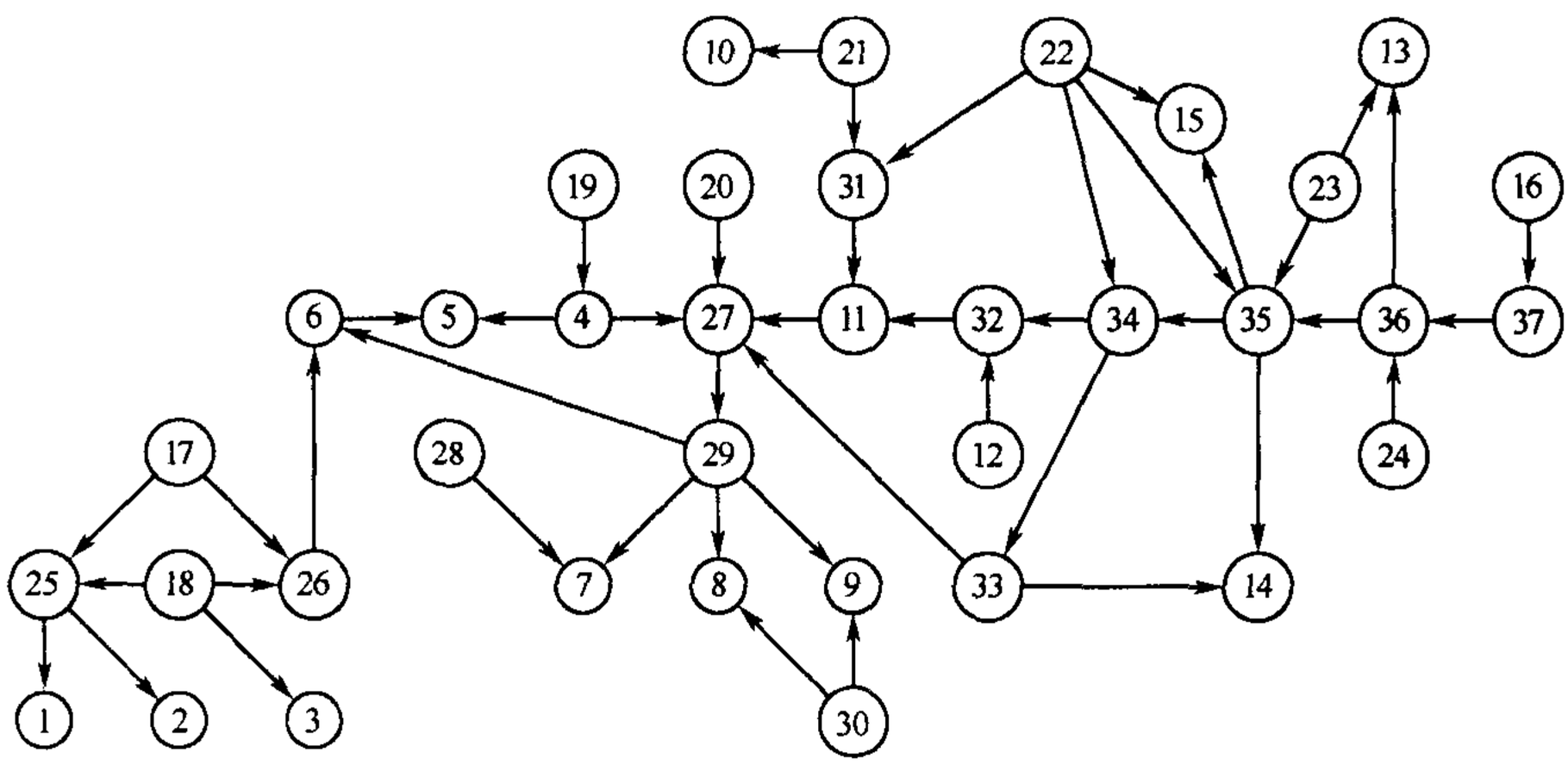


图 8.2　ALARM 网络

Cooper and Herskovits (1992) 从 ALARM 网中随机抽样出 10 000 个完整数据. 基于这些数据，利用 K2 算法进行结构学习. 在此过程中限定任一变量的父节点个数不超过 5 个，并选择了 ALARM 网络结构的一个拓扑序作为 K2 算法中的变量排序 ρ[①]. 然后，从所得的数据和变量排序出发，利用 K2 获得一个贝叶斯网结构. 这个过程在一台旧式 Macintosh-II 机器 上完成，共费时约 17 分钟. 所得的网络与 ALARM 网基本一致. 不同的是，少了边 "12→32"，多了边 "15→34". 因此可以说，在这个例子中，K2 找到了一个高质量的模型. 他们还用实验探索了样本量对 K2 的影响. 实验结果表明，在样本量是 3 000 时，K2 仍然可以获得同一个模型. 但是，当样本量进一步下降时，所得到的网络与原生网络 ALARM 的差距越来越大.

8.5.3 爬山法

爬山法的目标是要找出评分最高的模型. 它从一个初始模型出发开始搜索. 初始模型一般设为无边模型. 在搜索的每一步，它首先用**搜索算子** (search operator) 对**当前模型** (current model) 进行局部修改，得到一系列**候选模型** (candidate model)；然后计算每个候选模型的评分，并将最优候选模型与当前模型比较；若最优候选模型的评分大，则以它为下一个当前模型，继续搜索；否则，就停止搜索，并返回当前模型.

搜索算子有 3 个：**加边** (edge addition)、**减边** (edge deletion) 和 **转边** (edge reversal)，如图 8.3 所示. 加边即是在网络结构中添加一条边，减边即是减去一条边，转边是把一条边的方向翻转. 加边和转边算子的使用有一个前提，即不能在网络中形成有向圈.

爬山法可以使用任何评分函数. 不同的评分函数有不同的要求：CH 评分要求关于先验参数分布的超参数，而 HVL 及 CVL 评分则要求把数据分成训练数据和验证数据. 因此，需要处理的算法细节也有所不同. 图 8.4 给出的是使用罚项似然度的爬山法. 计算罚项似然度需要计算参数的最大似然估计. 所以，在图 8.4 中，我们把它视为模型 $\mathscr{G}$、模型参数 $\boldsymbol{\theta}$ 以及数据 $\mathscr{D}$ 的函数.

爬山法最费时的一步是计算候选模型 $\mathscr{G}'$ 的参数的最大似然估计 $\boldsymbol{\theta}'$ 及其评分 $f(\mathscr{G}', \boldsymbol{\theta}' \mid \mathscr{D})$. 由于 $\mathscr{G}'$ 与当前模型的差别不大，向量 $\boldsymbol{\theta}'$ 与当前模型参数的最大似然估计 $\boldsymbol{\theta}$ 相比，也只有几个维需要更新. 实际上，我们可以从 $\boldsymbol{\theta}$ 出发，利用式 (7.28)将受搜索算子影响的维重新计算，就可以得到 $\boldsymbol{\theta}'$. 另外，一些常见的罚项似然度，比方 BIC 和 AIC，都是可分解的，因此 $\mathscr{G}'$ 的评分可以从 $\mathscr{G}$ 的评分

---

① 所使用的变量顺序为 12 16 17 18 19 20 21 22 23 24 25 26 28 30 31 37 1 2 3 4 30 36 13 35 15 34 32 33 11 14 27 29 6 7 8 9 5.

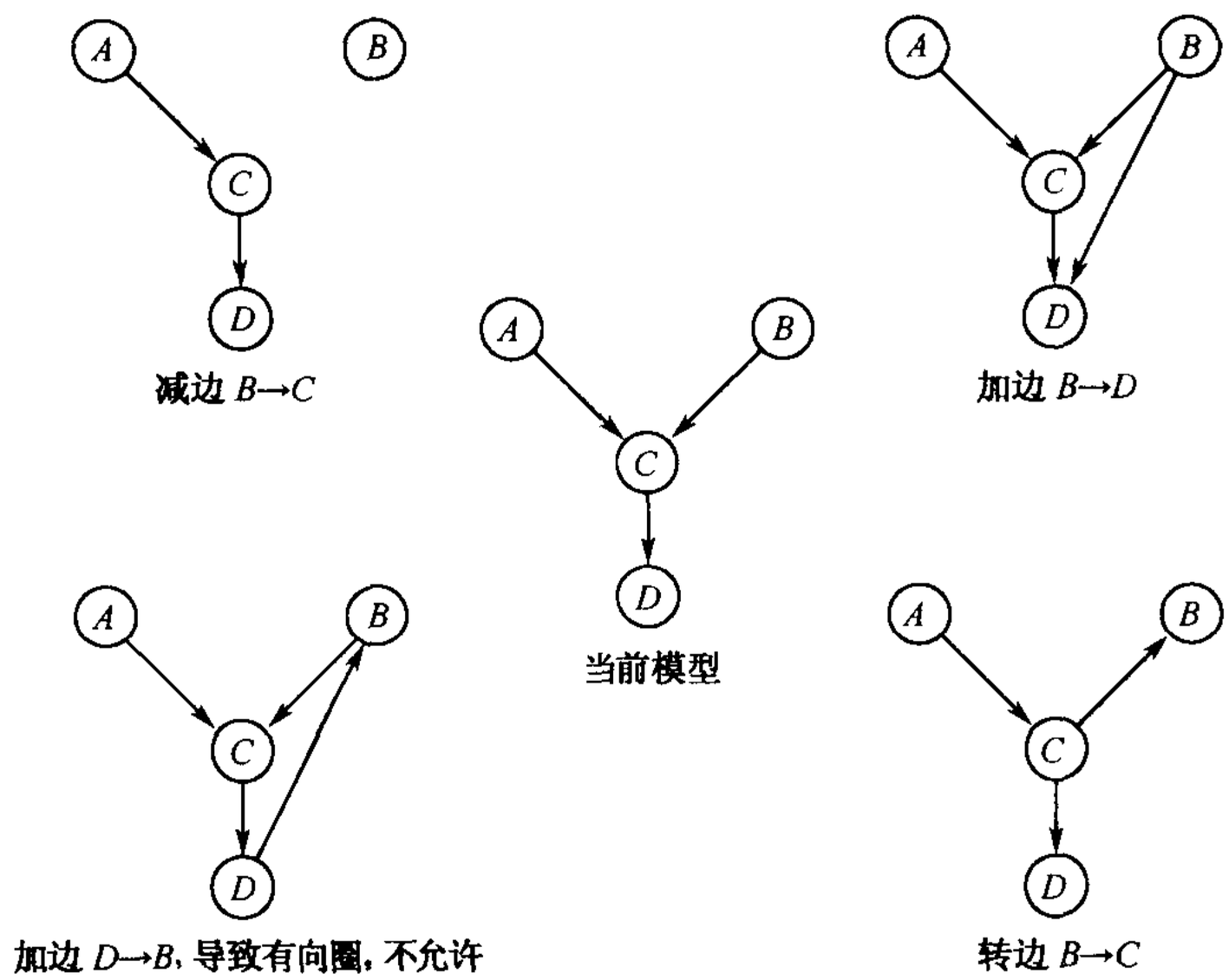


图 8.3 爬山法所用的 3 个搜索算子

LearnBN-HC($\mathbf{X}, \mathscr{D}, f, \mathscr{P}$)

输入：$\mathbf{X}$——一组变量；　$\mathscr{D}$——一组关于 $\mathbf{X}$ 的完整数据；
f——一个罚项似然度评分函数；　$\mathscr{G}^0$——一个初始贝叶斯网结构.

输出：一个贝叶斯网；

```
1: 𝒢←𝒢⁰; θ←𝒢 的参数的最大似然估计;
2: oldScore ←f(𝒢, θ|𝒟);
3: while (true)
4:     𝒢* ← null; θ* ← null; newScore ← −∞
5:     for(每个对 𝒢 做一次加边、减边或转边而得到的模型结构 𝒢′)
6:       θ′ ←𝒢′ 的参数的最大似然估计;
7:       tempScore ←f(𝒢′, θ′ |𝒟);
8:       if (tempScore > newScore)
9:         𝒢* ←𝒢′; θ* ← θ′ ; newScore ← tempScore;
10:      end if
11:    end for
12:    if (newScore > oldScore)
13:      𝒢←𝒢* ; θ←θ* ; oldScore ←newScore;
14:    else
15:      return(𝒢, θ)
16:    end if
17: end while
```

图 8.4 贝叶斯网学习爬山法

出发，通过简单计算获得. 例如，当使用 BIC 评分时，对于通过加边算子得到的候选模型 $\mathscr{G}'$，其评分可以通过式（8.46）获得.

爬山法有一个众所周知的缺点，即它可能因为陷入**局部最优**（local optimal）或是越不过**坪区**（plateau）而找不到**全局最优**（global maximum）. 克服此缺点的一个方法是多次运行爬山法，每次都从一个随机产生的新结构开始，最后取各次运行结果中最优的那个作为最后结果，这叫做**随机重复爬山法**（random restart hill-climbing）. 除此之外，也可以用组合优化领域中的其它启发式搜索方法，如**禁忌搜索**（tabu search）(Glover，1986，1989)、**模拟退火**（simulated annealing)(Kirkpatrick et al.，1983)，以及**遗传算法**（genetic algorithm)(Holland，1975；Goldberg，1989）等.

8.6 缺值数据结构学习

当数据有缺值时，仍可以使用爬山法 Learn BN-HC 来获得贝叶斯网模型. 但是，这时的参数估计（第 1 和 6 行）需要调用 EM 算法. 如果有 n 个变量，那么 LearnBN-HC 每一步产生 $O(n^2)$ 个候选模型，从而需要调用 EM 算法 $O(n^2)$ 次，因此计算的复杂度很高. 所以，在缺值数据情况下，LearnBN-HC 是不可行的. 本节介绍另一个方法，即**结构 EM 算法**（structure EM algorithm)，简称 **SEM** 算法. 它是针对一般的罚项似然度提出的. 为简洁起见，我们的介绍只针对 BIC 评分.

8.6.1 SEM 算法的基本思想

为理解 SEM 算法的基本思想，首先注意到 LearnBN-HC 在寻找 BIC 评分最高的模型结构的同时，也在计算其参数的最大似然估计. 换句话说，它对网络结构和参数同时进行优化. SEM 算法是 7.7 节中介绍的参数 EM 的一个自然推广，其基本思想是：从某初始模型结构 $\mathscr{G}^0$ 和参数 $\boldsymbol{\theta}^0$ 出发开始迭代，在进行了 t 次迭代得到了（$\mathscr{G}^t$，$\boldsymbol{\theta}^t$）后，第 $t+1$ 次迭代由以下两个步骤组成：

（1）基于（$\mathscr{G}^t$，$\boldsymbol{\theta}^t$）对数据进行修补，使之完整；

（2）基于修补后的完整数据 $\mathscr{D}^t$ 对模型及参数进行一步优化，得（$\mathscr{G}^{t+1}$，$\boldsymbol{\theta}^{t+1}$）.

这里有两点需要注意. 首先，在参数 EM 中，基于修补后完整数据 $\mathscr{D}^t$ 对参数进行一步优化得到的是基于 $\mathscr{D}^t$ 的最优参数，即最大似然估计. 但是，在 SEM 中，基于 $\mathscr{D}^t$ 进行一步优化未必得到基于 $\mathscr{D}^t$ 的最优模型结构，因为在完整数据下模型结构优化不是一步就可以完成的（参见第 8.5 节）. 第二，固定模型结构进行一步参数优化比进行一步结构优化要简单得多. 所以，SEM 不是每次迭代都

同时优化模型结构和参数，而是先固定模型结构，即规定 $\mathscr{G}^{t+1}=\mathscr{G}^t$，进行数次参数优化后，再进行一次结构加参数优化，如此交替进行.

8.6.2 SEM 算法

设（$\mathscr{G}$, $\boldsymbol{\theta}$）是一贝叶斯网，定义（$\mathscr{G}$, $\boldsymbol{\theta}$）的 **BIC 评分**如下①：

$$\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}\mid\mathscr{D})=\log P(\mathscr{D}\mid\mathscr{G},\boldsymbol{\theta})-\frac{d(\mathscr{G})}{2}\log m. \tag{8.47}$$

当 $\boldsymbol{\theta}$ 是最大似然估计 $\boldsymbol{\theta}^*$ 时，$\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}\mid\mathscr{D})$ 就是模型结构 $\mathscr{G}$ 的 BIC 评分：

$$\mathrm{BIC}(\mathscr{G}\mid\mathscr{D})=\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}^*\mid\mathscr{D}). \tag{8.48}$$

这一节假设数据有缺值，所关心的问题是在缺值数据情况下如何寻找 BIC 评分最高的贝叶斯网（$\mathscr{G}^*$, $\boldsymbol{\theta}^*$）：

$$\mathrm{BIC}(\mathscr{G}^*,\boldsymbol{\theta}^*\mid\mathscr{D})\geqslant\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}\mid\mathscr{D}),\forall\mathscr{G},\boldsymbol{\theta}. \tag{8.49}$$

设从某初始贝叶斯网出发，通过一系列优化步骤，SEM 得到了贝叶斯网（$\bar{\mathscr{G}}$, $\bar{\boldsymbol{\theta}}$）. 设 $\bar{\mathscr{D}}$ 是基于（$\bar{\mathscr{G}}$, $\bar{\boldsymbol{\theta}}$）对数据 $\mathscr{D}$ 进行修补而得到的完整数据. 与参数 EM 的情况类似式（7.47），一个贝叶斯网（$\mathscr{G}$, $\boldsymbol{\theta}$）的基于数据 $\bar{\mathscr{D}}$ 的 BIC 评分为

$$\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}\mid\bar{\mathscr{D}})=\sum_{l=1}^{m}\sum_{\mathbf{X}_l}P(\mathbf{X}_l\mid\mathbf{D}_l,\bar{\mathscr{G}},\bar{\boldsymbol{\theta}})\log P(\mathbf{D}_l,\mathbf{X}_l\mid\mathscr{G},\boldsymbol{\theta})-\frac{d(\mathscr{G})}{2}\log m, \tag{8.50}$$

其中 $\mathbf{X}_l$ 是 $\mathbf{D}_l$ 中所有值缺变量的集合. $\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}\mid\bar{\mathscr{D}})$ 也称为（$\mathscr{G}$, $\boldsymbol{\theta}$）的基于 $\mathscr{D}$ 的**期望 BIC 评分**（expected BIC score），可以记为 $Q(\mathscr{G},\boldsymbol{\theta}\mid\bar{\mathscr{G}},\bar{\boldsymbol{\theta}})$. 重复第 7.7.2 的推导，可得

$$Q(\mathscr{G},\boldsymbol{\theta}\mid\bar{\mathscr{G}},\bar{\boldsymbol{\theta}})=\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\bar{m}_{ijk}^{\mathscr{G}}\log\theta_{ijk}-\frac{d(\mathscr{G})}{2}\log m, \tag{8.51}$$

其中，

$$\bar{m}_{ijk}^{\mathscr{G}}=\sum_{l=1}^{m}P(X_i=k,\pi_{\mathscr{G}}(X_i)=j\mid\mathbf{D}_l,\bar{\mathscr{G}},\bar{\boldsymbol{\theta}}). \tag{8.52}$$

这里 $\pi_{\mathscr{G}}(X_i)$ 是 X_i 在 $\mathscr{G}$ 中的所有父节点的集合，$q_i=|\pi_{\mathscr{G}}(X_i)|$.

如果下一步是要优化参数，SEM 就规定 $\mathscr{G}=\bar{\mathscr{G}}$，并计算使 $Q(\bar{\mathscr{G}},\boldsymbol{\theta}\mid\bar{\mathscr{G}},\bar{\boldsymbol{\theta}})$ 达到最大的参数值 $\boldsymbol{\theta}'$. 在式（8.51）中，因为 $\mathscr{G}=\bar{\mathscr{G}}$，所以罚项为一常数. 根据第 1 章的推论 1.1，有

---

① 在给出算法 LearnBN-HC 时，已经提及罚项似然度可以看做是数据 $\mathscr{D}$、模型结构 $\mathscr{G}$ 以及模型参数的最大似然估计 $\boldsymbol{\theta}^*$ 三者的函数. 在那里，$\boldsymbol{\theta}^*$ 由 $\mathscr{D}$ 和 $\mathscr{G}$ 共同决定. 这里真正把 BIC 评分推广成 $\mathscr{D}$, $\mathscr{G}$, $\boldsymbol{\theta}$ 三者的函数.

$$\theta'_{ijk} = \frac{\overline{m}^{\mathscr{G}}_{ijk}}{\sum_{k=1}^{r_i} \overline{m}^{\mathscr{G}}_{ijk}}. \tag{8.53}$$

如果下一步是要同时优化模型结构和参数，SEM 首先构造出所有能够通过对 $\overline{\mathscr{G}}$ 进行一次加边、减边或转边而得到的候选模型，这些候选模型的集合记为 L；对其中任一候选模型 $\mathscr{G}$，SEM 用下式优化其参数，即

$$\theta_{ijk} = \frac{\overline{m}^{\mathscr{G}}_{ijk}}{\sum_{k=1}^{r_i} \overline{m}^{\mathscr{G}}_{ijk}}. \tag{8.54}$$

接着，利用式（8.51）求得 $Q(\mathscr{G}, \boldsymbol{\theta} \mid \overline{\mathscr{G}}, \overline{\boldsymbol{\theta}})$；最后再找出使 $Q(\mathscr{G}, \boldsymbol{\theta} \mid \overline{\mathscr{G}}, \overline{\boldsymbol{\theta}})$ 达到最大的模型，即

$$(\mathscr{G}', \boldsymbol{\theta}') = \arg\max_{\mathscr{G} \in L} \sup_{\boldsymbol{\theta}} Q(\mathscr{G}, \boldsymbol{\theta} \mid \overline{\mathscr{G}}, \overline{\boldsymbol{\theta}}). \tag{8.55}$$

SEM 的伪代码由图 8.5 给出．它先进行 R 次参数优化（第 2、3 行），接着同时优化模型结构及参数（第 5、6 行）．如果模型加参数优化得到评分更高的模型，SEM 就重复前面的运算，否则就停止并返回找到的 BIC 评分最高的贝叶斯网（第 7、8 行）．注意，一个贝叶斯网的 BIC 评分 $\mathrm{BIC}(\mathscr{G}, \boldsymbol{\theta} \mid \mathscr{D})$ 是根据式（8.47）计算的，与期望 BIC 评分不同．

SEM($\mathbf{X}, \mathscr{D}, \mathscr{G}^0, \boldsymbol{\theta}^{0,0}, R$)

输入：$\mathbf{X}$——一组变量；　$\mathscr{D}$——一组缺值数据；
$\mathscr{G}^0$——初始网络结构；　$\boldsymbol{\theta}^{0,0}$——初始参数值；
R——两次结构优化之间的参数优化次数；　δ——参数估计的收敛阈值．
输出：一个贝叶斯网；

```
1: for t=0 to ∞
2:     for r=0 to R-1
3:         θ^{t,r+1} ← arg sup_θ Q(𝒢^t, θ | 𝒢^t, θ^{t,r}) // 式(8.53)
4:     end for
5:     L ← 所有对 𝒢^t 做一次加边、减边或转边而得到的模型结构；
6:     (𝒢^{t+1}, θ^{t+1,0}) ← arg max_{𝒢∈L} sup_θ Q(𝒢, θ | 𝒢^t, θ^{t,R}); //式(8.54),式(8.55)
7:     if (BIC(𝒢^{t+1}, θ^{t+1,0} | 𝒟) ≤ BIC(𝒢^t, θ^{t,R} | 𝒟))
8:         return (𝒢^t, θ^{t,R})
9: end for
```

图 8.5　使用 BIC 评分的 SEM 算法

在缺值数据的情况下，SEM 的运算复杂度远远低于 LearnBN-HC 算法，主要原因是在优化模型结构时，SEM 不需要像 LearnBN-HC 那样调用 EM 算法来优化候选结构的参数．SEM 只需要优化最优候选结构的参数，即使这时，它也只进行 R 次参数优化，而不是像 EM 那样一直优化参数直到收敛为止．

下面谈一谈 SEM 的算法实现问题. 在第 7.7 节，我们讨论了怎样利用团树传播来实现参数 EM. SEM 的第 3 行所进行的正是参数 EM，所以可以用第 7.7 节描述的方法来实现.

团树传播也可以用来实现 SEM 的第 6 行. 这里对每一个候选结构 $\mathscr{G}$，首先需要计算其参数的、基于修补后数据 $\overline{\mathscr{D}}$ 的最大似然估计 $\boldsymbol{\theta}$，然后再计算 $Q(\mathscr{G},\boldsymbol{\theta}\mid\mathscr{G}^t,\boldsymbol{\theta}^{t,R})$. 根据式（8.51）、式（8.54）和式（8.52），我们必须对每一数据样本 $\mathbf{D}_l$ 和 每一变量 X_i 计算

$$P(X_i,\ \pi_{\mathscr{G}}(X_i)\mid\mathbf{D}_l,\mathscr{G}^t,\boldsymbol{\theta}^{t,R}). \tag{8.56}$$

为此，可以构造一个覆盖 $\mathscr{G}^t$ 的团树 $\mathscr{T}$，并用 $\boldsymbol{\theta}^{t,R}$ 将其初始化．对每个数据样本 $\mathbf{D}_l$，按照它设置证据，然后进行信息传播. 在信息传播结束后，就可以计算式（8.56）所示的后验概率分布. 具体地讲，因为我们在算法中只用到了加边、减边和转边三种搜索算子，所以对比 X_i 在 $\mathscr{G}^t$ 和 $\mathscr{G}$ 中的父节点，有以下两种情况：

（1）$\pi_{\mathscr{G}}(X_i)\subseteq\pi_{\mathscr{G}^t}(X_i)$，即 X_i 在 $\mathscr{G}$ 中的父节点在 $\mathscr{G}^t$ 中也是 X_i 的父节点. 这时，$\mathscr{T}$ 中必有一团包含 X_i 和 $\pi_{\mathscr{G}}(X_i)$. 于是 $P(X_i,\ \pi_{\mathscr{G}}(X_i)\mid\mathbf{D}_l,\ \mathscr{G}^t,\ \boldsymbol{\theta}^{t,R})$ 可以从该团中提取.

（2）$\exists X\in\pi_{\mathscr{G}}(X_i)$ s. t. $X\notin\pi_{\mathscr{G}^t}(X_i)$. 即 X_i 在 $\mathscr{G}$ 有一个父节点 X，它在 $\mathscr{G}^t$ 中不是 X_i 的父节点，但 X_i 在 $\mathscr{G}$ 中的其它父节点在 $\mathscr{G}^t$ 中都是 X_i 的父节点. 这时，在 $\mathscr{T}$ 中必有一个团 $\mathscr{C}$ 包含 X_i 和 $\pi_{\mathscr{G}}(X_i)\setminus\{X\}$. 另设 $\mathscr{C}'$ 是 $\mathscr{T}$ 中包含 X 并且距 $\mathscr{C}$ 最近的一个团. 那么，$P(X_i,\ \pi_{\mathscr{G}}(X_i)\mid\mathbf{D}_l,\ \mathscr{G}^t,\ \boldsymbol{\theta}^{t,R})$ 可以通过先重新从 $\mathscr{C}'$ 向 $\mathscr{C}$ 方向传递信息，然后从 $\mathscr{C}$ 处提取. 在重新从 $\mathscr{C}'$ 向 $\mathscr{C}$ 方向传递信息时，要注意不要消去变量 X.

在获得 $P(X_i,\ \pi_{\mathscr{G}}(X_i)\mid\mathbf{D}_l,\ \mathscr{G}^t,\ \boldsymbol{\theta}^{t,R})$ 之后，可以根据式（8.54）获得 $\mathscr{G}$ 的基于 $\overline{\mathscr{D}}$ 的参数的最大似然估计 $\boldsymbol{\theta}$，进而由式（8.51）求得 $Q(\mathscr{G},\ \boldsymbol{\theta}\mid\mathscr{G}^t,\ \boldsymbol{\theta}^{t,R})$ 并按照式（8.55）获得 $(\mathscr{G}^{t+1},\ \boldsymbol{\theta}^{t+1,0})$.

8.6.3 SEM 的收敛性

定理 8.3 设 $(\mathscr{G},\ \boldsymbol{\theta})$ 和 $(\overline{\mathscr{G}},\ \overline{\boldsymbol{\theta}})$ 为两个贝叶斯网. 如果 $Q(\mathscr{G},\ \boldsymbol{\theta}\mid\overline{\mathscr{G}},\ \overline{\boldsymbol{\theta}})\geqslant Q(\overline{\mathscr{G}},\ \overline{\boldsymbol{\theta}}\mid\overline{\mathscr{G}},\ \overline{\boldsymbol{\theta}})$，那么

$$\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}\mid\mathscr{D})\geqslant\mathrm{BIC}(\overline{\mathscr{G}},\overline{\boldsymbol{\theta}}\mid\mathscr{D}) \tag{8.57}$$

证明 与定理 7.4 的证明类似，有

$$\begin{aligned}\mathrm{BIC}(\mathscr{G},\boldsymbol{\theta}\mid\mathscr{D})&=Q(\mathscr{G},\boldsymbol{\theta}\mid\overline{\mathscr{G}},\overline{\boldsymbol{\theta}})\\&\quad-\sum_{l=1}^{m}\sum_{\mathbf{X}_l}P(\mathbf{X}_l\mid\mathbf{D}_l,\overline{\mathscr{G}},\overline{\boldsymbol{\theta}})\log P(\mathbf{X}_l\mid\mathbf{D}_l,\mathscr{G},\boldsymbol{\theta}).\end{aligned} \tag{8.58}$$

由于信息不等式（定理 1.3），有

$$\sum_{l=1}^{m}\sum_{\mathbf{X}_l}P(\mathbf{X}_l \mid \mathbf{D}_l,\bar{\mathscr{G}},\bar{\boldsymbol{\theta}})\log P(\mathbf{X}_l \mid \mathbf{D}_l,\bar{\mathscr{G}},\bar{\boldsymbol{\theta}})$$

$$\leqslant \sum_{l=1}^{m}\sum_{\mathbf{X}_l}P(\mathbf{X}_l \mid \mathbf{D}_l,\bar{\mathscr{G}},\bar{\boldsymbol{\theta}})\log P(\mathbf{X}_l \mid \mathscr{D},\mathscr{G},\boldsymbol{\theta}).$$

所以当 $Q(\mathscr{G},\ \boldsymbol{\theta} \mid \bar{\mathscr{G}},\ \bar{\boldsymbol{\theta}}) \geqslant Q(\bar{\mathscr{G}},\ \bar{\boldsymbol{\theta}} \mid \bar{\mathscr{G}},\ \bar{\boldsymbol{\theta}})$ 时，式（8.57）必然成立. 定理得证. □

考虑 SEM 在运行过程中所生成的贝叶斯网. 由于 $\boldsymbol{\theta}^{t,r+1}$ 的定义以及定理 8.3，有

$$\mathrm{BIC}(\mathscr{G}^t,\ \boldsymbol{\theta}^{t,r+1} \mid \mathscr{D}) \geqslant \mathrm{BIC}(\mathscr{G}^t,\ \boldsymbol{\theta}^{t,r} \mid \mathscr{D}). \tag{8.59}$$

再由算法的第 7 行，如果算法的运行未结束，就有

$$\mathrm{BIC}(\mathscr{G}^{t+1},\boldsymbol{\theta}^{t+1,0} \mid \mathscr{D}) \geqslant \mathrm{BIC}(\mathscr{G}^t,\ \boldsymbol{\theta}^{t,R} \mid \mathscr{D}). \tag{8.60}$$

所以有如下推论，

推论 8.1 无论是单独优化参数（第 3 行）还是同时优化模型结构和参数（第 6 行），SEM 所得的贝叶斯网的 BIC 评分总是单调上升的. 由于 BIC 评分有上界，所以 SEM 算法一定是收敛的. □

与爬山法一样，SEM 可能因为陷入局部最优或是越不过坪区而找不到全局最优. 同样，为克服此缺点，可以考虑使用随机重复、禁忌搜索以及模拟退火等启发式搜索方法.

8.7 文献介绍

关于模型选择的两本专著（Linhart and Zucchini，1986；Burnham and Anderson，2002）. 侧重于介绍 AIC 类型的准则. 想对 BIC 和 MDL 做深入了解的读者可以从 Lanterman 的综述文章（Lanterman，2001）开始.

BIC 评分 BIC（$\mathscr{G} \mid \mathscr{D}$）是对边缘似然度的一个近似. 它的近似误差是多少呢？Schwarz（1978）证明了在一定条件下，$\mathrm{BIC}(\mathscr{G} \mid \mathscr{D}) - \log P(\mathscr{D} \mid \mathscr{G}) = O(1)$，即是说其误差不随着样本量的增加而增加，因此在大样本时相对比较精确. 边缘似然度的拉普拉斯近似的误差，在一定条件下是 $\mathrm{Laplace}(\mathscr{D} \mid \mathscr{G}) - \log P(\mathscr{D} \mid \mathscr{G}) = O\left(\frac{1}{m}\right)$（Kass et al.，1988）. 除了拉普拉斯近似和 BIC 近似以外，也可以通过用蒙特卡洛方法直接近似计算式（8.14）中的积分来近似计算边缘似然度（Kass et al.，1988）. 蒙特卡洛法所得的结果可以比拉普拉斯近似和 BIC 近似更为精确，但它的计算复杂度也要高很多.

在第 8.4 节中提到，如果评分函数是一致的，那么当样本量趋于无穷时，原

则上可以重建数据的原生模型. 但是，具体怎样重建呢？这是近几年来贝叶斯网结构学习研究所关心的主要问题，相关结果参见有关文献（Chickering，2002，2003；Castelo and Kocka，2003）.

本章所介绍的是**基于评分的**（score-based）结构学习方法. 这类方法首先定义一个评分函数，然后寻找评分最高的结构. 除此之外还有另一类方法，称为**基于约束的**（constraint-based）结构学习方法. 这类方法从数据出发，首先做一些关于变量之间独立关系的假设检验，获得一些独立性关系，然后寻找与这些关系（约束）相容的网络结构. 作为一个例子，考虑3个变量 X，Y，Z. 设独立性检验发现它们之间的唯一独立关系是 $X \perp Y$，而另外的关系 $X \perp Z, Y \perp Z, X \perp Y \mid Z, X \perp Z \mid Y, Y \perp Z \mid X$ 全都不成立，那么，可能的贝叶斯网结构就只有一个，即“$X \rightarrow Z \leftarrow Y$”.

与基于评分的方法比较，基于约束的方法更为直观，有时得到的箭头可以理解为因果关系. 在速度方面，它也往往要快一些. 它的缺点主要是独立性检验比较容易出错，而一旦前面出错，就会影响后面的计算. 另外，它也不适用于缺值数据. 读者若想对基于约束的方法进行更深入的了解，可以参阅有关文献（Spirtes et al.，2000a；Spirtes and Meek，1995；Neapolitan，2003（第10章））.

第 9 章　隐结构模型学习

隐变量是取值未被观测到的变量. 本章所关心的问题是：在有隐变量的情况下，怎样通过数据分析获得贝叶斯网？这个问题十分困难，迄今还没有系统的研究. 本章介绍一些基本概念和初步结果，侧重讨论两种特殊的模型，即隐类模型和多层隐类模型. 与前几章不同，在本章中读者将会遇到许多悬而未决的问题.

9.1　隐变量与隐变量模型

一个变量是否被观测到是相对一组数据而言的. 设 X 为一变量，$\mathscr{D}$ 为一组数据. 如果 $\mathscr{D}$ 所有的样本都不包含 X 的取值，则称 X 相对于 $\mathscr{D}$ 是**隐变量** (latent variable). 为简洁起见，在讨论隐变量时往往不明确提及相应的数据组. 如果 $\mathscr{D}$ 中至少有一个样本包含 X 的取值，则称 X 为**显变量** (manifest variable) 或**观测变量** (observed variable). 从本章开始，我们用变量名 X 加上下标表示隐变量，用变量名 Y 加上下标表示显变量.

在应用中，隐变量可以分为两种，即值隐变量和势隐变量. 取值个数已知的隐变量称为**值隐变量** (value-hidden variable)，而取值个数未知的隐变量称为**势隐变量** (cardinality-hidden variable)①. 值隐变量虽然其取值未知，但取值个数是已知的. 势隐变量不仅其取值未被观测到，连取值个数也是未知的. 除非有特别说明，以后所提及的隐变量都是势隐变量.

包含至少一个隐变量的贝叶斯网模型称为**隐变量模型** (latent variable model). 在应用中，往往需要考虑多个隐变量，并且特别关心它们之间以及它们与显变量之间的关系. 在这样的情况下，隐变量模型也称为**隐结构模型** (latent structure model). 注意，“隐变量模型”是一个严格的数学概念，而“隐结构模型”则是一个描述性的概念.

隐结构模型学习 (latent structure model learning) 是要通过数据分析来决定模型以下几个方面的内容：

(1) 隐变量的个数；

(2) 以这些隐变量和显变量为节点的模型网络结构；

---

① 值隐变量和势隐变量的概念不是从英文翻译过来的，而是在这里首次提出的. 我们附上其英文翻译是为方便日后国际交流. 这种情况后面会出现多次，到时将不再一一指明.

(3) 每个隐变量的势；

(4) 模型参数.

当隐变量的个数以及模型的网络结构已知时，隐结构模型学习问题简化为**势学习**(cardinality learning) 问题. 进一步，当隐变量的势也已知时，势学习问题简化为参数学习问题，可以用 EM 算法来解决.

9.2　可分辨性及几个相关概念

设 $Y_1, Y_2, \cdots, Y_n$ 是一个隐变量模型 $\mathscr{G}$ 中的显变量. 如果有两个不同的参数值 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$，使得

$$P(Y_1, Y_2, \cdots, Y_n \mid \mathscr{G}, \boldsymbol{\theta}_1) = P(Y_1, Y_2, \cdots, Y_n \mid \mathscr{G}, \boldsymbol{\theta}_2), \tag{9.1}$$

则称模型 $\mathscr{G}$ 是**不可分辨的** (unidentifiable). 如果没有满足式 (9.1) 的不同的参数 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$，则称模型 $\mathscr{G}$ 是**可分辨的** (identifiable). 设 $\mathscr{D}$ 是一组关于 $Y_1, Y_2, \cdots, Y_n$ 的 i. i. d. 数据，如果式 (9.1) 成立，则有 $P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_1) = P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}_2)$. 这意味着基于数据 $\mathscr{D}$ 不可能分辨 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$.

与可分辨性密切相关的是有效维数的概念. 用 k 记 $|Y_1| \times |Y_2| \times \cdots \times |Y_n|$. 给定模型 $\mathscr{G}$ 的一个参数值 $\boldsymbol{\theta}$，联合分布 $P(Y_1, Y_2, \cdots, Y_n \mid \mathscr{G}, \boldsymbol{\theta})$ 是 R^{k-1} 空间中的一个点. 让 $\boldsymbol{\theta}$ 变动，所有这样的点形成一个**流形** (manifold). 这个流形称为模型 $\mathscr{G}$ 的**显象流形** (manifest image). 它的维数称为模型 $\mathscr{G}$ 的**有效维数** (effective dimension)，记为 $d_e(\mathscr{G})$. 另一方面，$\mathscr{G}$ 的独立参数的个数称为 $\mathscr{G}$ 的**标准维数**①，记为 $d(\mathscr{G})$. 如果 $\mathscr{G}$ 的标准维数大于其有效维数，则它是不可分辨的，标准维数与有效维数之差就是模型中不可分辨参数的个数.

一个模型不可分辨往往是因为它包含一些冗余部分. 这些冗余部分使得模型变得不必要地复杂. 此时应该将其简化，尽量去掉其冗余的部分，得到最简模型. 下面在隐变量模型的框架下把这个思想形式化.

设 $\mathscr{G}$ 和 $\mathscr{G}'$ 是两个具有相同显变量 $\mathbf{Y}=\{Y_1, Y_2, \cdots, Y_n\}$ 的隐变量模型. 我们称 $\mathscr{G}$ **显象包含** (manifestatively contains) $\mathscr{G}'$，如果对 $\mathscr{G}'$ 的任一参数值 $\boldsymbol{\theta}'$，$\mathscr{G}$ 有参数值 $\boldsymbol{\theta}$，使得

$$P(Y_1, Y_2, \cdots, Y_n \mid \mathscr{G}, \boldsymbol{\theta}) = P(Y_1, Y_2, \cdots, Y_n \mid \mathscr{G}', \boldsymbol{\theta}'). \tag{9.2}$$

换句话说，$\mathscr{G}$ 显象包含 $\mathscr{G}'$ 当且仅当 $\mathscr{G}$ 能够表示任何一个 $\mathscr{G}'$ 能够表示的关于 $\mathbf{Y}$ 的分布. 如果 $\mathscr{G}$ 与 $\mathscr{G}'$ 互相显象包含对方，则称 $\mathscr{G}$ 与 $\mathscr{G}'$ **显象分布等价** (manifesta-

① 设 $\mathscr{G}$ 的参数 $\boldsymbol{\theta}$ 是一个 l 维的向量，给定一组参数值，就得到 R^l 空间的一个点. 让参数值变化，就得到 R^l 空间的一个流形. 这个流形的维数正好是模型的独立参数的个数，即 $\mathscr{G}$ 的标准维数.

tively equivalent in distribution). 显象分布等价的模型具有相同的有效维数. 如果它们有相同的标准维数，则称它们**显象等价** (marginally equivalent). 在不与第 8.4 节中所定义的等价概念混淆的情况下，显象等价可以简称为**等价** (equivalent). 如果不存在与 $\mathscr{G}$ 显象分布等价且标准维数又比 $\mathscr{G}$ 低的模型，那么称模型 $\mathscr{G}$ 是**最简的** (parsimonious). "最简模型" 的概念不具有操作性，一个模型是否最简，往往难以判断. 后面将引入一个具有操作性的概念来对它进行近似，即正则性 (regularity) 的概念. 最后需要注意的是，最简模型也可能是不可分辨的. 换句话说，有些不可分辨性是无法去掉的.

9.3 隐变量模型选择

最大似然估计的唯一性是我们在第 8.3 节中推导边缘似然度的拉普拉斯近似时用到的重要条件. 然而，在第 9.4.2 节将会看到，这一条件在有隐变量的时候不成立. 于是，BIC 评分不一定适用于隐变量模型选择. 有鉴于此，Geiger et al. (1996) 在隐变量模型的显象流形上对边缘似然度的大样本近似重新进行了推导，得到一个新的评分准则，称为 **BICe 评分**：

$$\mathrm{BICe}(\mathscr{G} \mid \mathscr{D}) = \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*) - \frac{d_e(\mathscr{G})}{2}\log m, \tag{9.3}$$

其中 $\mathscr{D}$ 是一组数据，m 是 $\mathscr{D}$ 中的样本个数，$\mathscr{G}$ 是一个隐变量模型，而 $\boldsymbol{\theta}^*$ 是 $\mathscr{G}$ 的参数的一个最大似然估计，$d_e(\mathscr{G})$ 是 $\mathscr{G}$ 的有效维数. BICe 评分与 BIC 评分的唯一区别是，前者的罚项使用有效维数，而后者的罚项使用标准维数.

另一个评分准则是由 Cheeseman and Stutz (1995) 给出的，因此称为 **CS 评分**. 设 $\mathscr{D}'$ 是基于 $(\mathscr{G}, \boldsymbol{\theta}^*)$ 将 $\mathscr{D}$ 进行修补而得到的完整数据. 推导 CS 评分的出发点为

$$\log P(\mathscr{D} \mid \mathscr{G}) = \log\left[P(\mathscr{D}' \mid \mathscr{G}) \frac{\int P(\mathscr{D}, \boldsymbol{\theta} \mid \mathscr{G})\mathrm{d}\boldsymbol{\theta}}{\int P(\mathscr{D}', \boldsymbol{\theta} \mid \mathscr{G})\mathrm{d}\boldsymbol{\theta}}\right]. \tag{9.4}$$

由于 $\mathscr{D}'$ 是完整数据，上式右端的第一项 $\log P(\mathscr{D}' \mid \mathscr{G})$ 是基于 $\mathscr{D}'$ 的 CH 评分，可以用式 (8.21) 来计算. 至于分数部分，对分子分母的对数同时在 $\boldsymbol{\theta}^*$ 附近进行大样本近似，得

$$\begin{aligned}\mathrm{CS}(\mathscr{G} \mid \mathscr{D}) = {} & \log P(\mathscr{D}' \mid \mathscr{G}) + \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*) - \frac{d_e(\mathscr{G})}{2}\log m \\ & - \log P(\mathscr{D}' \mid \mathscr{G}, \boldsymbol{\theta}^*) + \frac{d(\mathscr{G})}{2}\log m.\end{aligned} \tag{9.5}$$

由于近似是对分子和分母同时进行的，因此近似误差会部分地相互抵消.

注意，式（9.5）中既有模型 $\mathscr{G}$ 的有效维数 $d_e(\mathscr{G})$，又有其标准维数 $d(\mathscr{G})$. 这是因为相对于 $\mathscr{D}$，$\mathscr{G}$ 有隐变量；但是相对于 $\mathscr{D}'$，$\mathscr{G}$ 却没有隐变量. Cheeseman 和 Stutz（1995）并没有考虑到这一点. 他们给出的评分其实为

$$\mathrm{CS}(\mathscr{G} \mid \mathscr{D}) = \log P(\mathscr{D}' \mid \mathscr{G}) + \log P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*) - \log P(\mathscr{D}' \mid \mathscr{G}, \boldsymbol{\theta}^*). \quad (9.6)$$

而式（9.5）是由 Chickering 和 Heckerman（1997）给出的. 在有必要区分式（9.5）和式（9.6）所给出的评分时，将前者称为**修正 CS 评分**，后者称为**原始 CS 评分**.

尽管 BICe 评分和 CS 评分在理论上优于 BIC 评分，但是迄今为止还没有系统的实验数据来验证这一点. 我们在对多层隐变量模型的研究中得到的经验表明，BIC 评分和原始 CS 评分的表现是基本一致的，它们都优于 AIC 评分和 HVL 评分（见第 9.6.4 节）.

使用 BICe 评分和修正 CS 评分有一个困难，即对模型有效维数的计算. 对于某些特殊情况，有一些分解和近似方法（Kocka and Zhang，2002；Zhang and Kocka，2004b；Chen et al.，2005）. 但是在一般情况下，需要计算高维矩阵的秩（Geiger et al.，1996），计算复杂度很高. 所以，虽然 Chickering 和 Heckerman（1997）提出了修正 CS 评分，但在实验中还是假设 $d_e(\mathscr{G}) = d(\mathscr{G})$ 以简化计算.

总而言之，隐变量模型选择有许多问题尚未解决，需要进一步系统地研究. 在找到更好的评分函数之前，一般先暂时使用 BIC 评分.

9.4　隐类模型

隐类模型（latent class model），简称 **LC 模型**，是由一个隐变量和多个显变量所组成的贝叶斯网，如图 9.1 所示. 在应用中，隐变量 X 代表未被观测到的隐因子，如学生的“分析能力”；显变量 $Y_1, Y_2, \cdots, Y_n$ 代表隐因子的外在表现，如学生的“数学成绩”、“物理成绩”以及“化学成绩”等. 变量 X 的不同取值对应不同的类，这些类未被观测到，因而称为**隐类**（latent class）. 如果 X 代表的是“分析能力”，那么隐类就是“分析能力高（之学生）”、“分析能力低（之

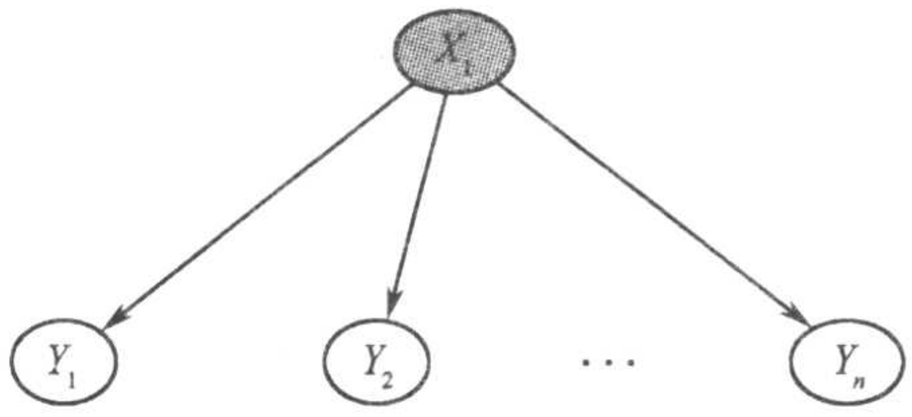


图 9.1　隐类模型图示，其中 X 是隐变量，$Y_1, Y_2, \cdots, Y_n$ 是显变量

学生)”等等. 隐类模型与朴素贝叶斯模型具有相同的网络结构. 不同的是，后者中的类变量是显变量，而前者中的类变量是隐变量.

9.4.1 正则性

正如第 9.2 节所提及的那样，应尽量除去隐变量模型的不可分辨部分，以得到最简模型. 为了方便判断一个隐类模型是否最简，我们引入正则性的概念. 一个隐类模型是**正则的**（regular)，如果其隐变量的势 $|X|$ 和其显变量的势 $|Y_i|$ $(i=1,2,\cdots,n)$ 满足如下关系：

$$|X| \leqslant \frac{\prod_{i=1}^{n} |Y_i|}{\max_{i=1}^{n} |Y_i|}. \tag{9.7}$$

定理 9.1 如果一个隐类模型 $\mathscr{G}$ 不是正则的，那么就存在另一个与 $\mathscr{G}$ 显象分布等价的隐类模型 $\mathscr{G}'$，其标准维数低于 $\mathscr{G}$ 的标准维数.

证明 不失一般性，设 $Y_n=\max_{i=1}^{n} |Y_i|$. 因为 $\mathscr{G}$ 非正则，有

$$|X| > \prod_{i=1}^{n-1} |Y_i|.$$

设 $\mathscr{G}'$ 是由隐变量 Z 和显变量 $Y_i(i=1,2,\cdots,n)$ 所组成的隐类模型，其中

$$|Z| = \prod_{i=1}^{n-1} |Y_i|. \tag{9.8}$$

显然，$\mathscr{G}$ 的标准维数小于 $\mathscr{G}$ 的标准维数. 下面证明 $\mathscr{G}'$ 与 $\mathscr{G}$ 显象等价.

由于 $\mathscr{G}$ 与 $\mathscr{G}'$ 的唯一区别是 $\mathscr{G}$ 中的隐变量 X 比 $\mathscr{G}'$ 中的隐变量 Z 有更多的状态，因此 $\mathscr{G}$ 显像包含 $\mathscr{G}'$. 剩下需要证明的是，$\mathscr{G}'$ 显像包含 $\mathscr{G}$. 我们证明一个更强的结果：$\mathscr{G}'$ 能够表示任何一个 关于显变量 Y_1，Y_2，…，Y_n 的联合分布 $P(Y_1, Y_2, \cdots, Y_n)$.

由于式 (9.8)，Z 的状态的个数正好是 $Y_i(i=1,2,\cdots,n-1)$ 的所有状态组合的个数. 于是，可以在 Ω_Z 和 $\Omega_{Y_1}\times\Omega_{Y_2}\times\cdots\times\Omega_{Y_{n-1}}$ 之间建立起一一对应，从而可以把 Z 的状态写成 $(y_1, y_2, \cdots, y_{n-1})$，其中 y_i 是 Y_i 的一个状态. $\mathscr{G}'$ 的参数 $\boldsymbol{\theta}'$ 由概率分布 $P(Z \mid \mathscr{G}', \boldsymbol{\theta}')$ 和 $P(Y_i \mid Z, \mathscr{G}', \boldsymbol{\theta}')$ $(i=1, 2, \cdots, n)$ 组成. 分别规定它们如下：

$$P(Z=(y_1, \cdots, y_{n-1}) \mid \mathscr{G}', \boldsymbol{\theta}') = \sum_{Y_n} P(Y_1=y_1,\cdots,Y_{n-1}=y_{n-1},Y_n),$$

$$P(Y_i \mid Z=(y_1, \cdots, y_{n-1}), \mathscr{G}', \boldsymbol{\theta}') = \begin{cases} 1, \text{当 } Y_i=y_i(i=1,2,\cdots,n-1) \\ 0, \text{其它情况} \end{cases},$$

$$P(Y_n \mid Z=(y_1, \cdots, y_{n-1}), \mathscr{G}', \boldsymbol{\theta}') = \frac{P(Y_1=y_1,\cdots,Y_{n-1}=y_{n-1},Y_n)}{\sum_{Y_n} P(Y_1=y_1,\cdots,Y_{n-1}=y_{n-1},Y_n)}.$$

由于以上三式，有

$$P(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n \mid \mathscr{G}', \boldsymbol{\theta}') = P(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n).$$

即 $\mathscr{G}'$ 可以表示 $P(Y_1, Y_2, \cdots, Y_n)$. 定理得证. □

定理 9.1 告诉我们，非正则隐类模型不是最简的隐类模型. 这是因为 $\mathscr{G}$ 与 $\mathscr{G}'$ 显象分布等价，它们的有效维数相同，而 $\mathscr{G}$ 的标准维数大于 $\mathscr{G}'$ 的标准 维数. 定理 9.1 的证明则告诉我们怎样将一个非正则模型**正则化**（regularization），即减少隐变量的状态个数.

下面我们将只考虑正则模型. 需要注意的是，虽然在正则化的过程中去掉了模型中的某些不可分辨部分，但是得到的正则隐类模型不一定是可分辨的①. Kocka and Zhang（2002）给出了几个正则隐类模型的有效维数和标准维数，如表 9.1 所示. 表中的隐类模型用隐变量和显变量的势所组成的序列来表示. 例如，2：3，3 表示一个具有 2 个显变量的隐类模型，其中隐变量的势为 2，两个显变量的势都是 3.

表 9.1　几个正则隐类模型的有效维数和标准维数

模型	有效维数	标准维数
2：2，2	3	5
3：2，2，2	7	11
4：2，2，2	7	15
2：3，3	7	9
3：4，5	17	23
4：3，3，3	25	27

我们看到，尽管表中的所有隐类模型都是正则的，但它们的标准维数都大于它们的有效维数，说明它们是不可分辨的.

9.4.2　隐变量模型选择与大样本理论

在第 9.3 节讨论模型选择时曾经提到，在隐变量模型中，最大似然估计是不唯一的. 所以在对边缘似然度进行大样本近似时，拉普拉斯方法并不适用. 下面以隐类模型为例说明隐变量模型中最大似然估计的不唯一性.

---

① Goodman（1974）提出隐类模型可分辨的一个必要条件是 $|X| \leqslant \dfrac{\prod_{i=1}^{n} |Y_i|}{\sum_{i=1}^{n} |Y_i| - n + 1}$. 显然这个条件比式（9.7）更强.

设 $\mathscr{G}$ 是一个由隐变量 X 和显变量 $Y_1, Y_2, \cdots, Y_n$ 所组成的隐类模型，$\mathscr{D}$ 是关于显变量的一组数据，其中 X 的取值缺失. 设 $\boldsymbol{\theta}^*$ 是 $\mathscr{G}$ 的参数的一个最大似然估计. 另外，设 x_1 和 x_2 为 X 的两个不同的状态，并且 $P(X=x_1 \mid \mathscr{G}, \boldsymbol{\theta}^*) \neq P(X=x_2 \mid \mathscr{G}, \boldsymbol{\theta}^*)$. 定义另一组参数值 $\boldsymbol{\theta}'$ 为

$$P(X=x \mid \mathscr{G}, \boldsymbol{\theta}') = \begin{cases} P(X=x_2 \mid \mathscr{G}, \boldsymbol{\theta}^*), \text{当 } x=x_1 \\ P(X=x_1 \mid \mathscr{G}, \boldsymbol{\theta}^*), \text{当 } x=x_2 \\ P(X=x \mid \mathscr{G}, \boldsymbol{\theta}^*), \text{其它情况} \end{cases} \tag{9.9}$$

$$P(Y_i \mid X=x, \mathscr{G}, \boldsymbol{\theta}') = \begin{cases} P(Y_i \mid X=x_2, \mathscr{G}, \boldsymbol{\theta}^*), \text{当 } x=x_1 \\ P(Y_i \mid X=x_1, \mathscr{G}, \boldsymbol{\theta}^*), \text{当 } x=x_2 \\ P(Y_i \mid X=x, \mathscr{G}, \boldsymbol{\theta}^*), \text{其它情况} \end{cases} \tag{9.10}$$

那么有

$$P(Y_1, Y_2, \cdots, Y_n \mid \mathscr{G}, \boldsymbol{\theta}^*) = P(Y_1, Y_2, \cdots, Y_n \mid \mathscr{G}, \boldsymbol{\theta}'). \tag{9.11}$$

这意味着

$$P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}^*) = P(\mathscr{D} \mid \mathscr{G}, \boldsymbol{\theta}'). \tag{9.12}$$

于是隐类模型中最大似然估计不唯一.

上面揭示的事实可以推广为：在隐变量模型中，调换隐变量状态的名称（renaming states of latent variables）不会改变显变量的分布. 这是一种特殊的不可分辨性，称为**隐状态的不可分辨性**（unidentifiability of latent states）.

9.4.3 隐类模型学习算法

隐类模型学习又称为**隐类分析**（latent class analysis）. 它的出发点是一组关于显变量 $Y_1, Y_2, \cdots, Y_n$ 的数据. 通过对这些数据进行分析，要确定的是：

（1）隐变量 X 的势；

（2）概率分布 $P(X)$ 及 $P(Y_i \mid X)$ $(i=1, 2, \cdots, n)$.

确定隐变量的势等于是确定隐类的个数，确定 $P(X)$ 及 $P(Y_i \mid X)$ 就是确定每个隐类的统计特性. 因此，隐类分析是一种**基于模型的聚类**（model-based clustering）分析.

图 9.2 给出了一个学习隐类模型的爬山算法 LearnLCM-HC. 其中LCM($\mathbf{Y}$, C) 代表由显变量集合 $\mathbf{Y}$ 和一个势为 C 的隐变量所组成的隐类模型. 另外，尽管诸如 BICe 、CS、BIC 和 AIC 的评分函数在理论上只依赖模型结构 $\mathscr{G}$ 和数据 $\mathscr{D}$，但是在实际中需要计算 $\mathscr{G}$ 的参数的最大似然估计，然后基于这个估计计算模型的评分，所以在 LearnLCM-HC 算法中，评分函数 f 被视为是模型 $\mathscr{G}$、数据 $\mathscr{D}$ 以及参数估计 $\boldsymbol{\theta}$ 三者的函数.

LearnLCM-HC 算法首先考虑隐变量的势为 2 的情况，调用 EM 估计模型的参数，并计算模型的评分. 然后逐步增加隐变量的势，直到模型评分不

再增加为止.

LearnLCM-HC($\mathbf{Y}, \mathscr{D}, f, \delta$)

输入：$\mathbf{Y}$——一组显变量；　$\mathscr{D}$——一组关于 $\mathbf{Y}$ 的数据；
　　f——一个模型评分函数；　δ—— EM 算法收敛阈值.

输出：一个隐类模型；

```
C←2;
𝒢←LCM(Y,C);
θ←EM(𝒢,𝒟,δ);
oldScore←f(𝒢,θ|𝒟);
while (true)
    C←C+1;
    𝒢'←LCM(Y,C);
    θ'←EM(𝒢',𝒟,δ);
    newScore←f(𝒢',θ'|𝒟)
    if(newScore > oldScore)
        𝒢←𝒢'; θ←θ'; oldScore←newScore;
    else
        return (𝒢,θ);
    end if
end while
```

图 9.2　隐类模型学习算法

9.4.4　仿真试验

第 9.3 节介绍了两个隐变量模型选择标准，即 BICe 评分和 CS 评分. 用它们来指导隐类模型学习的效果如何呢？为了回答这个问题，我们进行了一些仿真实验. 我们首先随机产生隐类模型，从中抽样得到关于显变量的数据；然后从这些数据出发，基于不同的评分函数，利用算法 LearnLCM-HC 获得隐类模型；最后比较原生模型中隐变量的势与学习结果中隐变量的势.

实验一共选用了 100 个原生模型，每个模型中显变量的个数在 3 与 10 之间，显变量的势在 2 与 5 之间，而隐变量的势在 2 和 10 之间. 原生模型要求是正则的，否则，首先将其正则化. 实验考虑了 4 个不同的样本量，即 500、1000、5000 和10000. 作为比较对象，实验还涉及了另外 2 个评分函数，即 BIC 评分和 AIC 评分.

实验结果用散点图（scatter plot）来总结，分别见图 9.3～9.6. 图中的每一点代表一次实验，它的横坐标表示原生模型中隐变量的势，简称为**真实势**. 而它的纵坐标表示学习结果中隐变量的势，简称为**估计势**. 不同次实验可能对应同一个点. 一个点旁边的数字是与该点对应的实验的总数. 对每一评分函数和样本量的组合，有一个散点图，于是共有 16 幅. 这些图按样本量被分为 4 组.

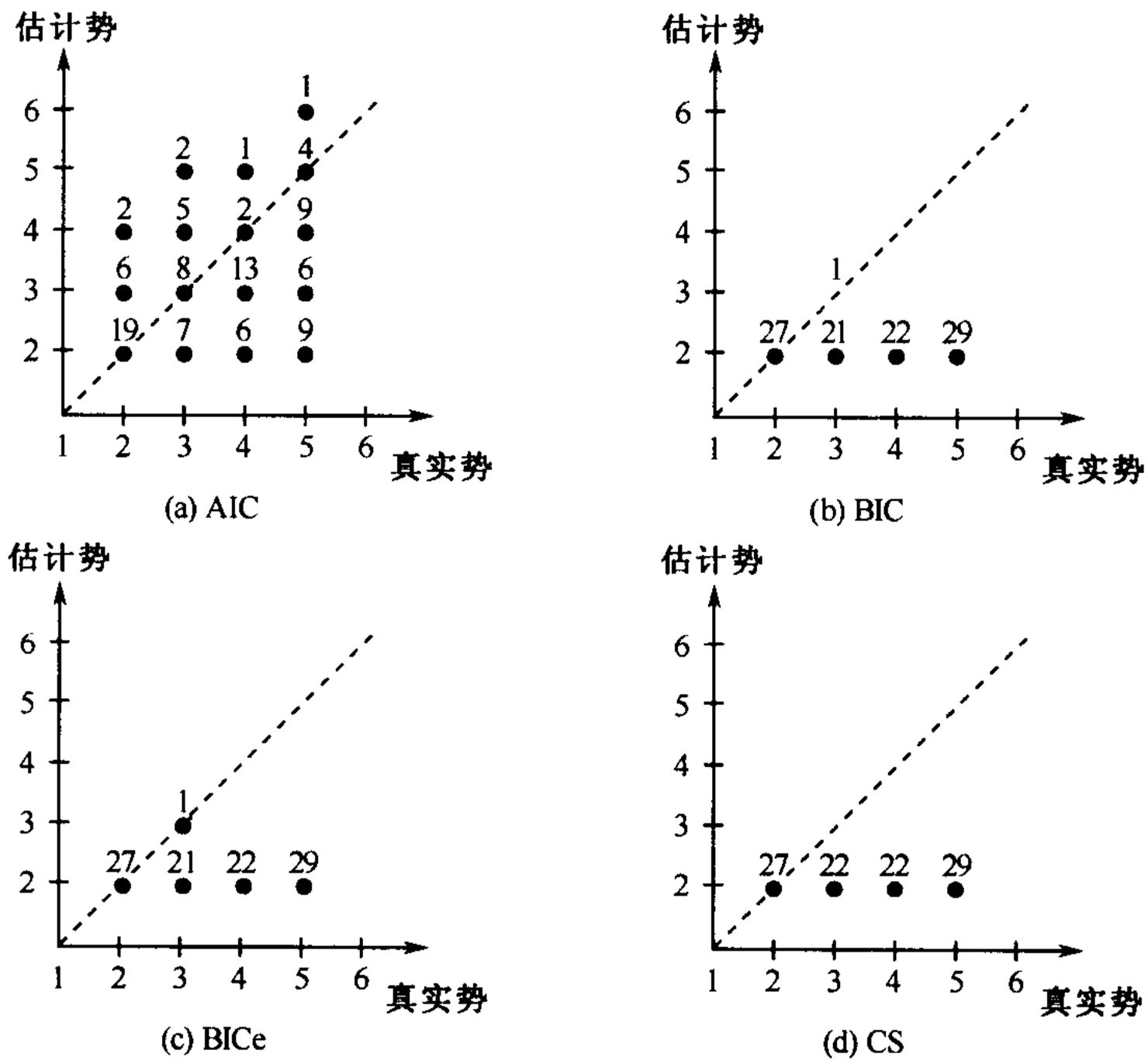


图 9.3　样本量为 500 时算法 LearnLCM-HC 使用各评分函数的学习结果

(图中点的横坐标为真实势，即原生模型中隐变量的势；纵坐标表示估计势，即学习结果中隐变量的势．一个点旁边的数字是与该点对应的实验的总数．图 9.4～图 9.6 的解释相同)

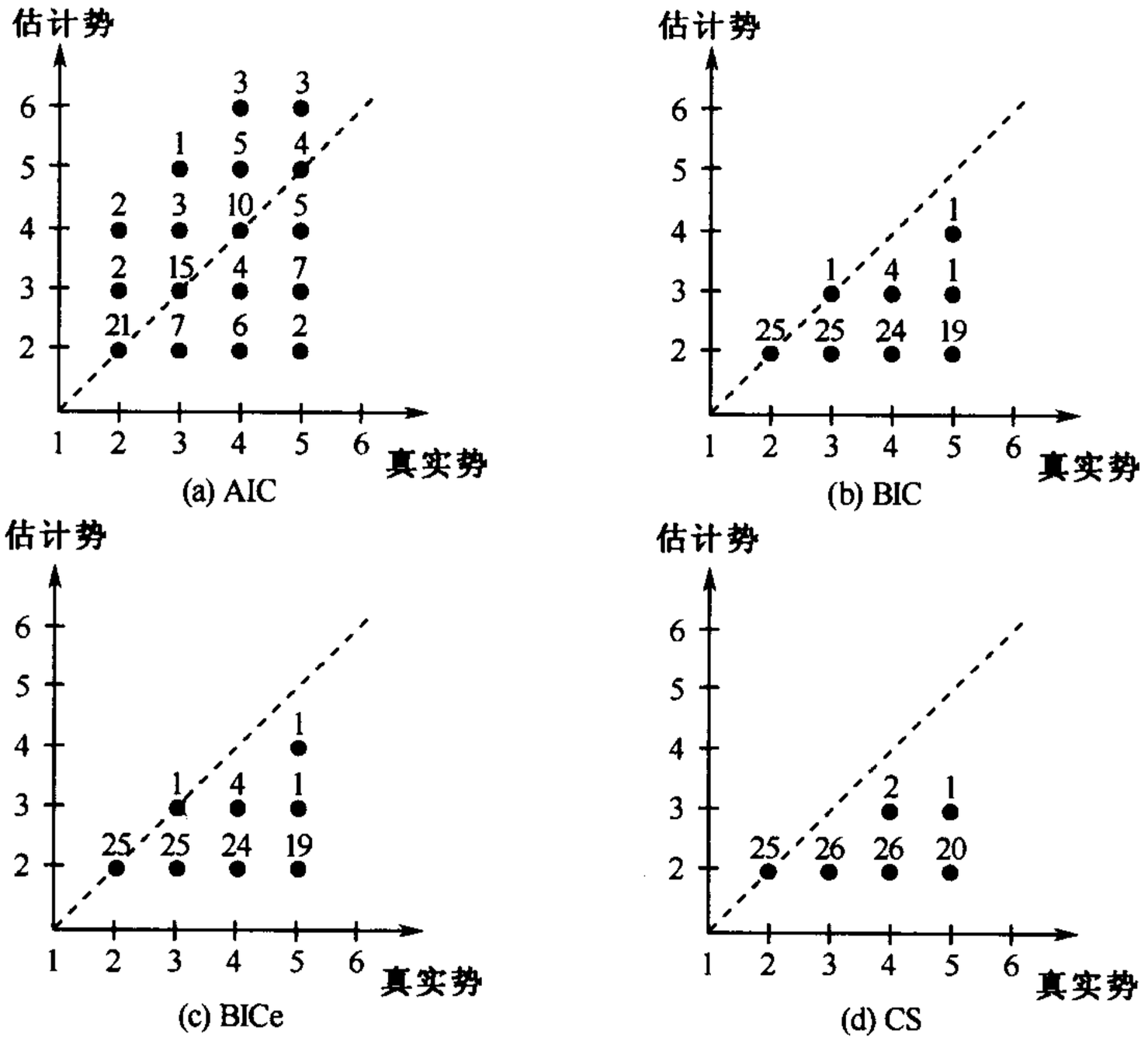


图 9.4　样本量为 1000 时算法 LearnLCM-HC 使用各评分函数的学习结果

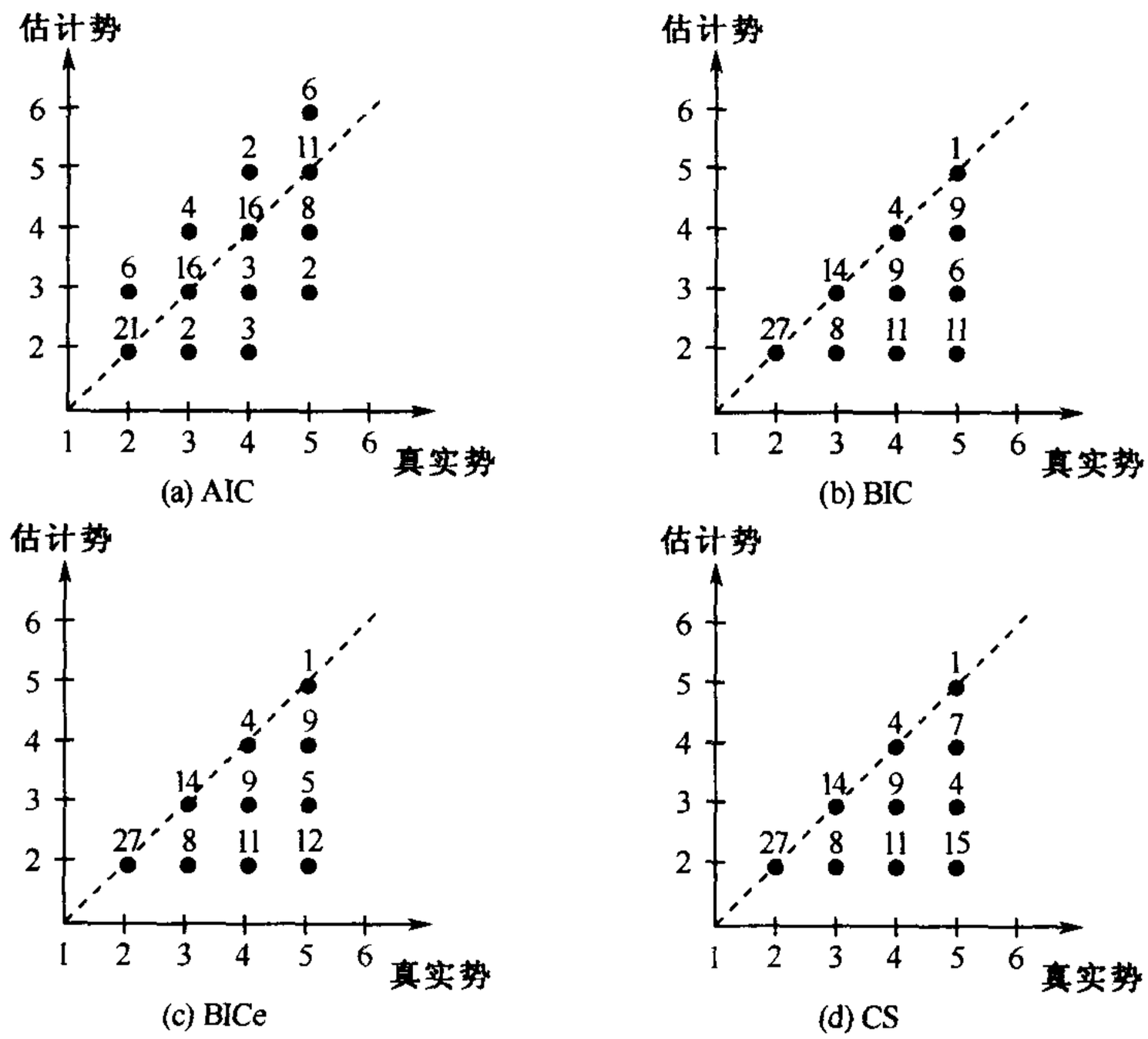


图 9.5　样本量为 5000 时算法 LearnLCM-HC 使用各评分函数的学习结果

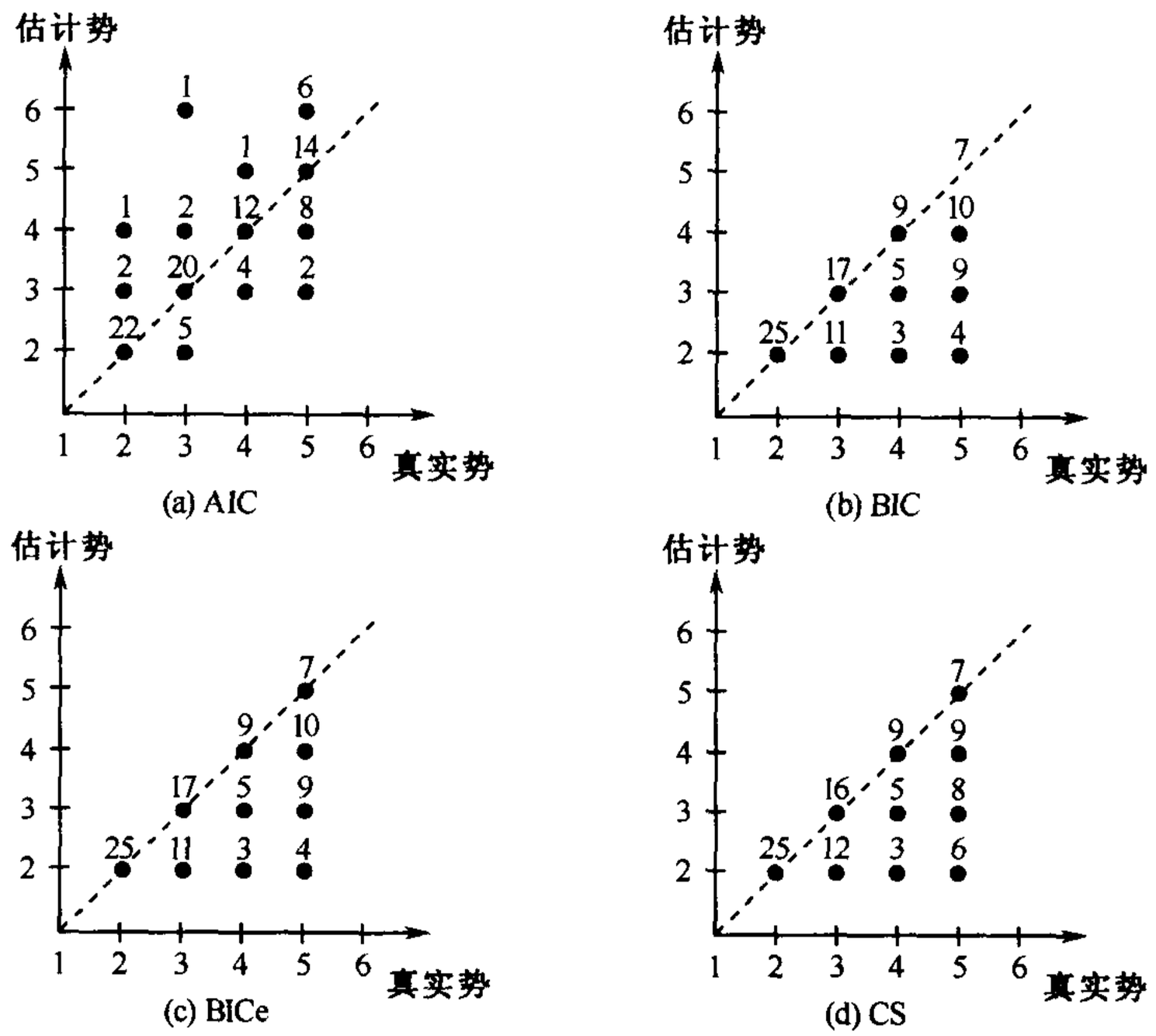


图 9.6　样本量为 10000 时算法 LearnLCM-HC 使用各评分函数的学习结果

实验结果表明，利用 BIC 和 BICe 所得的结果几乎完全一样，它们与利用 CS 所得的结果基本一致. 所以，就隐类模型学习而言，BIC，BICe 和 CS 之间没有明显区别. 这与 Chickering 和 Heckerman（1997）的结论不同，其主要原因是实验所涉及的样本量和显变量个数不同. 他们的实验涉及的样本量是 50、100、200 和 400，涉及的显变量个数是 32、64 和 128.

学习得到的隐变量的势是对原生模型中隐变量的真实势的估计. 利用 BIC，BICe 和 CS 所得的估计是保守的，估计势不大于真实势. 随着样本量的增加，估计势与真实势相等的次数越来越多，这表明估计质量越来越好①. 另一方面，利用 AIC 所得的估计不是保守的：估计势既有大于真实势的时候，也有小于真实势的时候. 且当样本量增加时，估计质量无明显改进.

9.5 多层隐类模型

多层隐类模型（hierarchical latent class model），简称 **HLC 模型**，是一种特殊的隐变量模型，其中：

（1）网络结构是一棵非平凡有根树②，其中根节点至少有两个子节点；

（2）所有叶节点代表的都是显变量，因此也称为**显节点**；所有内节点代表的都是隐变量，因此也称为**隐节点**.

图 9.7 给出了一个 HLC 模型，其中 X_1，X_2，X_3 是隐变量，而 Y_1，Y_2，…，Y_7 是显变量. 显然，LC 模型是 HLC 模型的一个特例，它只含一个隐变量.

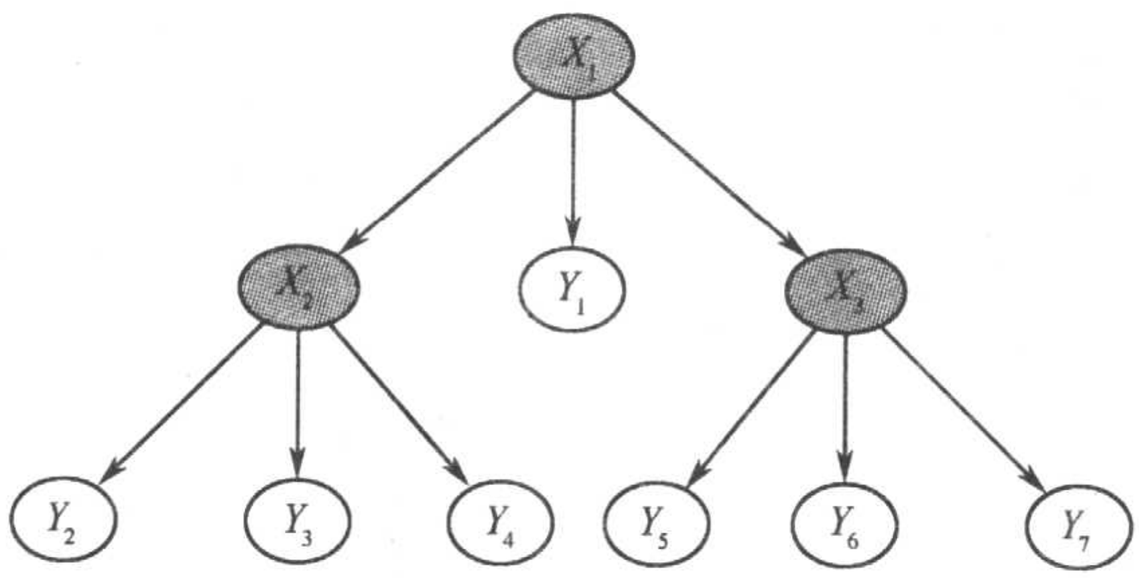


图 9.7 一个 HLC 模型：X_1，X_2，X_3 是隐变量，Y_1，Y_2，…，Y_7 是显变量

① 注意，即使在样本量是 10000 的情况，估计势往往还是低于真实势，这与原生模型中的参数设置有关. 对于隐变量 X 的两个不同的取值 x_1 和 x_2，如果 $P(Y_i \mid X=x_1)$ 与 $P(Y_i \mid X=x_2)$ 相近，那么基于数据难以区分这两个隐类，从而容易将它们并为一类.

② 在图论中，树是一个不含圈且连通的无向图. 这里的有根树指的是底图为树，并且有且仅有一个根节点的有向图.

9.5.1　走根运算与模型等价

设 X_1 是一个 HLC 模型 $\mathscr{G}_1$ 的根节点，X_2 是 X_1 的一个子节点. 如果 X_2 是隐变量，那么可以转换边“$X_1 \rightarrow X_2$”的方向而得到另一个 HLC 模型 $\mathscr{G}_2$，如图 9.8 所示. 此时，模型的根从 X_1 走到了 X_2，因此从 $\mathscr{G}_1$ 到 $\mathscr{G}_2$ 的过程称为**走根**(root walking).

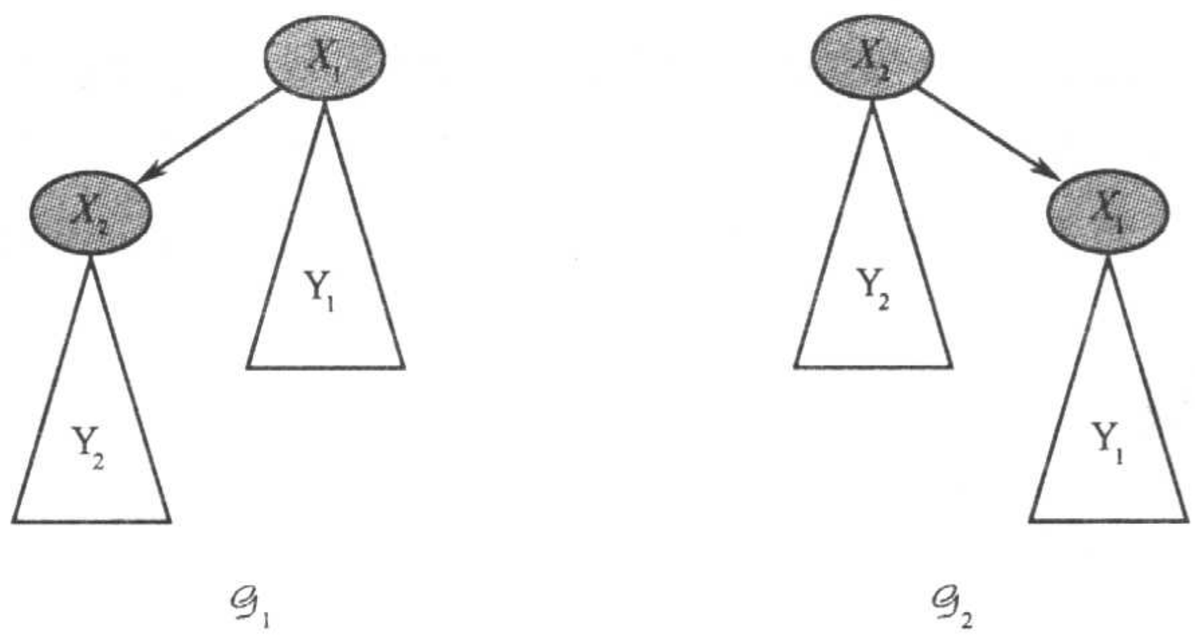


图 9.8　走根运算模型 $\mathscr{G}_1$ 的根从 X_1 走到 X_2 得 $\mathscr{G}_2$

定理 9.2　设 $\mathscr{G}_1$ 和 $\mathscr{G}_2$ 是两个 HLC 模型. 如果 $\mathscr{G}_2$ 可以从 $\mathscr{G}_1$ 出发经过一次或多次走根而得到，那么 $\mathscr{G}_1$ 和 $\mathscr{G}_2$ 显象等价.

证明　不失一般性，设 $\mathscr{G}_2$ 是从 $\mathscr{G}_1$ 经过一次走根得到的. 首先证明 $\mathscr{G}_1$ 和 $\mathscr{G}_2$ 显象分布等价. 设 $\mathscr{G}_1$ 和 $\mathscr{G}_2$ 的根分别为 X_1 和 X_2，并分别用 $\mathbf{Y}$ 和 $\mathbf{Z}$ 记模型中所有显变量和所有变量的集合. 对于 $\mathscr{G}_1$ 的任一参数值 $\boldsymbol{\theta}_1$，定义 $\mathscr{G}_2$ 的参数值 $\boldsymbol{\theta}_2$ 为

$$P(X_2 \mid \mathscr{G}_2, \boldsymbol{\theta}_2) = \sum_{X_1} P(X_1 \mid \mathscr{G}_1, \boldsymbol{\theta}_1) P(X_2 \mid X_1, \mathscr{G}_1, \boldsymbol{\theta}_1), \tag{9.13}$$

$$P(X_1 \mid X_2, \mathscr{G}_2, \boldsymbol{\theta}_2) = \frac{P(X_1 \mid \mathscr{G}_1, \boldsymbol{\theta}_1) P(X_2 \mid X_1, \mathscr{G}_1, \boldsymbol{\theta}_1)^{①}}{P(X_2 \mid \mathscr{G}_2, \boldsymbol{\theta}_2)}, \tag{9.14}$$

$$P(Z \mid \pi(Z), \mathscr{G}_2, \boldsymbol{\theta}_2) = P(Z \mid \pi(Z), \mathscr{G}_1, \boldsymbol{\theta}_1), \forall Z \in \mathbf{Z} \text{ 且 } Z \neq X_1, X_2. \tag{9.15}$$

不难看出，

$$P(\mathbf{Z} \mid \mathscr{G}_1, \boldsymbol{\theta}_1) = P(\mathbf{Z} \mid \mathscr{G}_2, \boldsymbol{\theta}_2), \tag{9.16}$$

从而

$$P(\mathbf{Y} \mid \mathscr{G}_1, \boldsymbol{\theta}_1) = P(\mathbf{Y} \mid \mathscr{G}_2, \boldsymbol{\theta}_2). \tag{9.17}$$

所以，$\mathscr{G}_2$ 显象包含 $\mathscr{G}_1$. 由于对称性，$\mathscr{G}_1$ 也显象包含 $\mathscr{G}_2$. 于是，$\mathscr{G}_1$ 和 $\mathscr{G}_2$ 显象分布等价.

---

① 为了简洁，我们假设 $P(X_2 \mid \mathscr{G}_2, \boldsymbol{\theta}_2) > 0$.

另一方面，除了 X_1 和 X_2 的概率分布以外，$\mathscr{G}_1$ 和 $\mathscr{G}_2$ 的其它参数完全一样. 而 $\{P(X_1 \mid \mathscr{G}_1, \boldsymbol{\theta}_1), P(X_2 \mid X_1, \mathscr{G}_1, \boldsymbol{\theta}_1)\}$ 与 $\{P(X_2 \mid \mathscr{G}_2, \boldsymbol{\theta}_2), P(X_1 \mid X_2, \mathscr{G}_2, \boldsymbol{\theta}_2)\}$ 均蕴含 $|X_1| \, |X_2| - 1$ 个独立参数. 因此 $\mathscr{G}_1$ 与 $\mathscr{G}_2$ 具有相同的标准维数，所以它们显象等价. 定理得证. □

在图 9.7 所示的 HLC 模型中，如果模型的根从 X_1 走到 X_2，就得到图 9.9(a) 中的模型；如果模型的根走到 X_3，就得到图 9.9(b) 中的模型. 根据定理 9.2，这两个模型与图 9.7 所示的模型显象等价. 一般地讲，HLC 模型的根走到任何一个隐节点都得到一个与原模型显象等价的模型. 这就意味着不可能通过数据分析决定 HLC 模型的根. 换句话说，HLC 模型的根是不可分辨的. 这一特殊的不可分辨性称为 HLC 模型的**根不可分辨性**（unidentifiability of root).

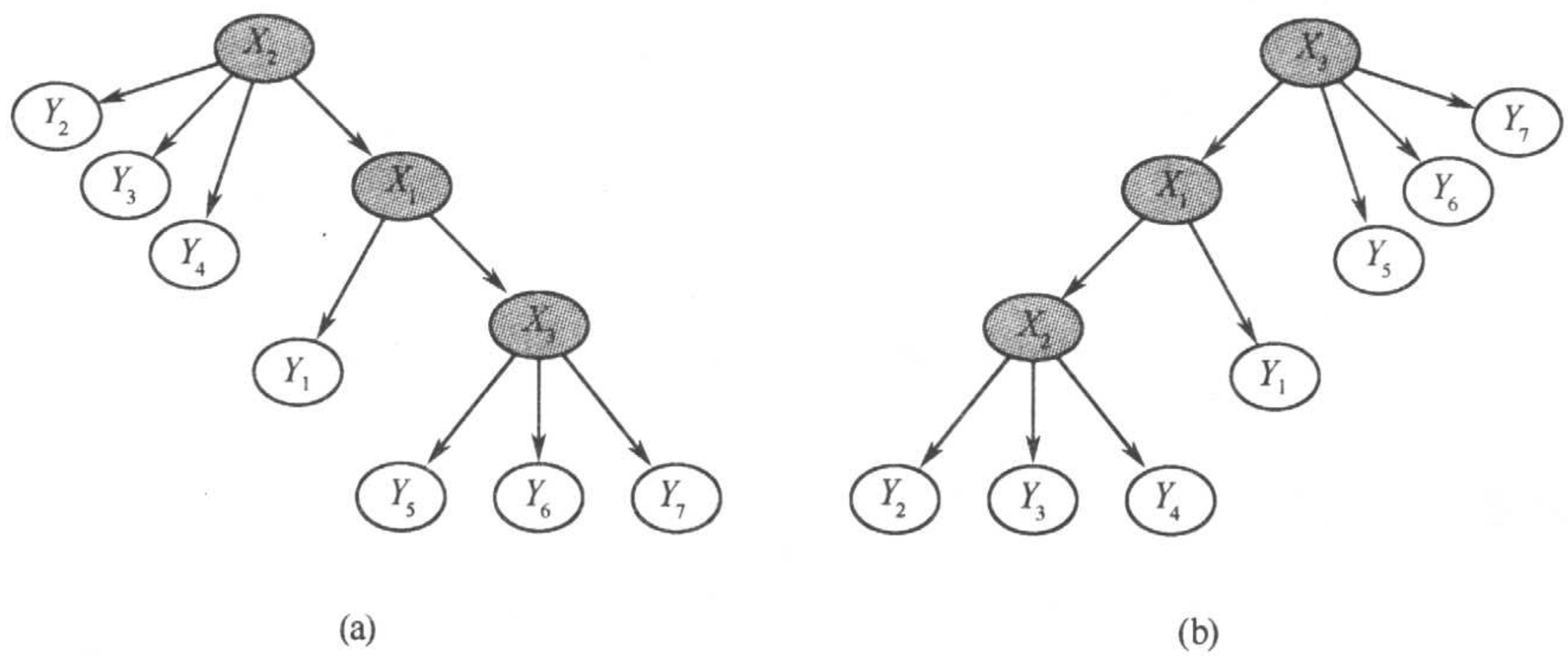


图 9.9 走根运算示例

9.5.2 无根 HLC 模型

由于根不可分辨，通过数据分析只可能得到无根 HLC 模型. 在语义上，**无根 HLC 模型**是一个树上的**马尔可夫随机场**（Markov random field)：它的节点表示随机变量，节点间的边代表变量间的直接依赖关系；它的定量内容包括：每一个节点 Z 附有一个非负函数 $f(Z)$，刻画该节点的“势位”；每一条边“Z_i-Z_j”附有一个非负函数 $f(Z_i, Z_j)$，刻画变量 Z_i 和 Z_j 之间的相互影响. 这些函数称为**势位函数**（potential).

贝叶斯网表示的是一个联合概率分布的分解，而马尔可夫随机场也表示一个联合分布的分解. 实际上，将场中所有势位函数相乘，然后将其结果归一化，就得到场中所有变量的联合分布. 关于马尔可夫随机场的一般介绍请读者参见 (Kindermann and Snell，1980).

对于一个 HLC 模型，设它所表示的联合概率分布为 $P(\mathbf{Z})$，它的根节点为

X，则可以按如下方法构造一个无根 HLC 模型：对根节点，将其势位函数设为 $P(X)$；对其它节点，将其势位函数设为单位函数 **1**；对每一条边“$Z_i \rightarrow Z_j$”，去掉其方向并将势位函数设为 $P(Z_j \mid Z_i)$. 这个无根 HLC 模型所表示的联合分布就是 $P(\mathbf{Z})$. 从图 9.7 所示的模型出发，按上述方法所得到的无根 HLC 模型的结构如图 9.10 所示.

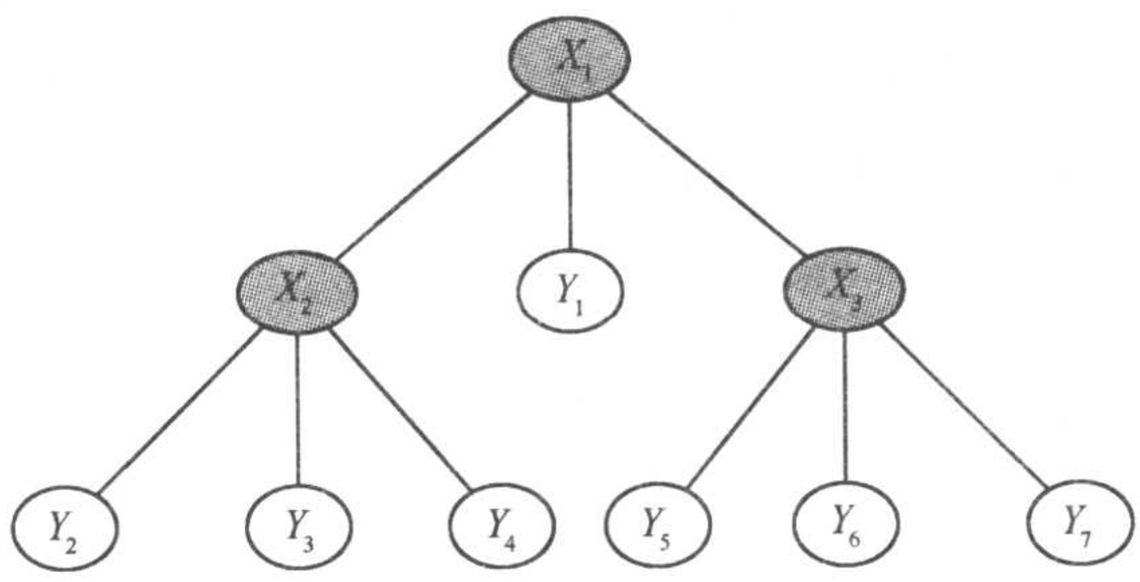


图 9.10　与图 9.7 所示模型相对应的无根 HLC 模型

给定一个无根 HLC 模型，设它所表示的联合概率分布为 $P(\mathbf{Z})$，可以按如下方法构造一个 HLC 模型：任选一个内节点作为根节点，并相应地设置边的方向；将每一个节点 Z 的条件分布设为 $P(Z|\pi(Z))$. 这个 HLC 模型所表示的联合概率分布也是 $P(\mathbf{Z})$. 由于根节点选择的任意性，可以得到很多这样的 HLC 模型，这些模型可以通过走根运算而互达，所以它们都是显象等价的. 因此，一个无根 HLC 模型所代表的是一个 HLC 模型等价类.

既然模型的根是不可分辨的，（有根）HLC 模型的概念是否还有意义？回答是肯定的. 这有两个原因：首先，在有些应用中，模型的根可以在模型的诠释阶段依据背景知识来确定，而且根的确定是必要的. 读者将会在第 10 章看到这样的一个例子. 其次，（有根）HLC 是特殊的贝叶斯网，因此，一些关于贝叶斯网的 算法可以直接用来处理（有根）HLC 模型，这为算法实现带来了方便.

除非特别说明，以后所提及的 HLC 模型均指无根 HLC 模型. 一个无根 HLC 模型的标准维数定义为其所对应的任一有根 HLC 模型的标准维数，而不可分辨性、有效维数、显象包含、显象分布等价以及显象等价等概念对无根 HLC 模型同样适用.

9.5.3　正则性

设 X 为 HLC 模型中的一个隐变量，用 $Z_1, Z_2, \cdots, Z_k$ 记所有与 X 相邻的节点. 一个 HLC 模型称为是**正则的**（regular）如果：

（1）它至少有两个显变量；

(2) 对任一隐变量 X，有

$$|X| \leqslant \frac{\prod_{i=1}^{k} |Z_i|}{\max_{i=1}^{k} |Z_i|}, \tag{9.18}$$

而且当 $k=2$ 且 X 的一个相邻节点是隐变量时，上面的不等式严格成立.

定理 9.3 设 $\mathscr{G}$ 是一个 HLC 模型. 如果 $\mathscr{G}$ 不是正则的，那么一定存在一个与 $\mathscr{G}$ 显象分布等价的 HLC 模型 $\mathscr{G}'$，其标准维数低于 $\mathscr{G}$ 的标准维数.

证明 假设 $\mathscr{G}$ 不是正则的，那么有两种情况：

(1) $\mathscr{G}$ 中存在隐变量 X，使得

$$|X| > \frac{\prod_{i=1}^{k} |Z_i|}{\max_{i=1}^{k} |Z_i|}; \tag{9.19}$$

(2) $\mathscr{G}$ 中存在隐变量 X，它有两个相邻节点，且其中之一是隐节点，并有

$$|X| \geqslant \frac{\prod_{i=1}^{k} |Z_i|}{\max_{i=1}^{k} |Z_i|}. \tag{9.20}$$

不失一般性，设 $Z_k = \max_{i=1}^{k} |Z_i|$.

先考虑第一种情况. 构造另一 HLC 模型 $\mathscr{G}'$，它与 $\mathscr{G}$ 一样，只是隐变量 X 被另一隐变量 X' 取代，且

$$|X'| = \prod_{i=1}^{k-1} |Z_i|. \tag{9.21}$$

很明显，$\mathscr{G}'$ 的标准维数低于 $\mathscr{G}$ 的标准维数，而且 $\mathscr{G}$ 显象包含 $\mathscr{G}'$. 下面证明 $\mathscr{G}'$ 也显象包含 $\mathscr{G}$.

设 $\boldsymbol{\theta}$ 为 $\mathscr{G}$ 的任一参数值，其中包括 X 的势位函数以及 X 与其相邻节点 Z_1，$Z_2, \cdots, Z_k$ 之间的边的势位函数. 用 $f(X, Z_1, Z_2, \cdots, Z_k)$ 记这些函数的乘积，并定义 $g(Z_1, Z_2, \cdots, Z_k) = \sum_X f(X, Z_1, Z_2, \cdots, Z_k)$. 由于式 (9.21)，可以在 $\Omega_{X'}$ 与 $\Omega_{Z_1} \times \Omega_{Z_2} \times \cdots \times \Omega_{Z_{k-1}}$ 之间建立起一一对应，因此 X' 的状态可以写成 $(z_1, z_2, \cdots, z_{k-1})$，其中 z_i 是 Z_i 的一个状态. 定义 $\mathscr{G}'$ 的参数值 $\boldsymbol{\theta}'$ 定义如下：设 X' 的势位函数 $f(X')$ 为单位函数 **1**；设边 "X'-Z_i" 的势位函数为

$$f_i(X' = (z_1, z_2, \cdots, z_{k-1}), Z_i) = \begin{cases} 1, \text{若 } Z_i = z_i \\ 0, \text{若否} \end{cases}, i = 1, 2, \cdots, k-1, \tag{9.22}$$

$$f_k(X' = (z_1, z_2, \cdots, z_{k-1}), Z_k) = g(Z_1 = z_1, \cdots, Z_{k-1} = z_{k-1}, Z_k). \tag{9.23}$$

不难看出，

$$\sum_{X'} f(X') \prod_{i=1}^{k} f_i(X', Z_i) = g(Z_1, Z_2, \cdots, Z_k).$$

设 $\mathscr{G}'$ 的其它节点及边的势位函数与 $\mathscr{G}$ 中相同，从而有

$$P(\mathbf{Y} \mid \mathscr{G}, \boldsymbol{\theta}) = P(\mathbf{Y} \mid \mathscr{G}', \boldsymbol{\theta}'),$$

其中 $\mathbf{Y}$ 是模型中所有显变量的集合．这就证明了 $\mathscr{G}'$ 显象包含 $\mathscr{G}$．从而证明了在式（9.19）成立时定理成立．

下面考虑第二种情况．这时 X 只有两个相邻节点，即 Z_1 和 Z_2，且它们之一是隐节点．不失一般性，设 Z_1 是隐节点．构造模型 $\mathscr{G}'$ 如下：除去 X，并将 Z_1 和 Z_2 直接相连．为了计算 $\mathscr{G}'$ 与 $\mathscr{G}$ 的标准维数的差别，想象把它们的根都定在 Z_1．不难看出，

$$\begin{aligned}
d(\mathscr{G}') - d(\mathscr{G}) = & \mid Z_1 \mid - 1 + \mid Z_1 \mid (\mid Z_2 \mid - 1) \\
& - [\mid Z_1 \mid - 1 + \mid Z_1 \mid (\mid X \mid - 1) + \mid X \mid (\mid Z_2 \mid - 1) \\
= & \mid Z_1 \mid \mid Z_2 \mid - \mid X \mid (\mid Z_1 \mid + \mid Z_2 \mid - 1) \\
& < \mid Z_1 \mid \mid Z_2 \mid - \mid X \mid \max\{\mid Z_1 \mid, \mid Z_2 \mid\} (\text{因为} \mid Z_1 \mid, \mid Z_2 \mid \geqslant 2) \\
& \leqslant 0, \quad (\text{因为式}(9.20))
\end{aligned}$$

即 $\mathscr{G}'$ 的标准维数低于 $\mathscr{G}$ 的标准维数．

最后需要证明 $\mathscr{G}$ 与 $\mathscr{G}'$ 显象分布等价．我们把这作为一个练习留给读者．定理证毕． □

定理 9.3 告诉我们，非正则 HLC 模型一定不是最简的．而其证明则告诉我们怎样将一个非正则模型正则化，那就是按照式（9.21）减少隐变量的状态个数，或者消去一个隐变量．

9.5.4　正则模型空间的有限性

在一个 HLC 模型中，可以加任意多的隐节点，使其结果仍是 HLC 模型．所以，对一组给定的显变量，有无穷多可能的 HLC 模型．本节证明，在排除非正则模型后，剩下的正则模型的个数是有穷的．

首先证明几个引理．一个 HLC 模型由其网络结构、显变量的势以及隐变量的势组成．如果隐变量的势未给出，则称其为**预 HLC 模型**（pre-HLC model）．对一个预 HLC 模型，如果给出其中每个隐变量的势，就得到一个 HLC 模型．显然，一个预 HLC 模型可以生成无穷多个 HLC 模型．

引理 9.1　一个预 HLC 模型最多只能生成有穷个正则 HLC 模型．

证明　用 K 记所有显变量的势的乘积．在正则 HLC 模型中，任意一个隐变量的势不可能超过 K．如果一个预 HLC 模型有 h 个隐变量，那么它所生成的正则模型的个数不可能超过 K^h，因此是有穷的．引理证毕． □

在一个正则 HLC 模型中，每个隐节点至少有两个相邻节点. 如果一个隐节点刚好只有两个相邻节点，则称它是**单连的**（singly connected).

引理 9.2 在正则 HLC 模型中，两个单连隐节点不可能是相邻的.

证明 用反证法. 设一个正则 HLC 模型 $\mathscr{G}$ 中有两个相邻的单连隐节点 Z_1 和 Z_2. 根据正则性的定义，必有 $|Z_1|<|Z_2|$ 和 $|Z_2|<|Z_1|$. 这是一个矛盾，所以引理成立. 证毕. □

一个预 HLC 模型称为**正则的**如果它至少包含两个显节点，且没有相邻的单连隐节点. 如果其中完全没有单连隐节点，则称它是**严格正则的**. 根据引理 9.2，非正则预 HLC 模型不可能生成任何正则 HLC 模型. 下面证明，正则预 HLC 模型的个数是有穷的.

引理 9.3 对于一个具有 n 个显变量的预 HLC 模型 $\mathscr{G}$，有

(1) 如果 $\mathscr{G}$ 是正则的，那么 $\mathscr{G}$ 中的隐节点个数不超过 $3n$；

(2) 如果 $\mathscr{G}$ 是严格正则的，那么 $\mathscr{G}$ 中的隐节点个数不超过 n.

证明 先证 (2). 用 h 记 $\mathscr{G}$ 中隐变量的个数. $\mathscr{G}$ 中节点总数为 $n+h$，因此边的数目为 $n+h-1$. 另一方面，每个显节点只在一条边中出现，而由于严格正则性，每个隐节点至少在 3 条不同的边中出现. 由于每条边与两个节点相关联，所以 $\mathscr{G}$ 中至少有 $(n+3h)/2$ 条边. 于是 $n+h-1\geqslant(n+3h)/2$，从而$h\leqslant n-2$.

接下来证明 (1). 在 $\mathscr{G}$ 中消去所有的单连隐节点，得到另一个预 HLC 模型 $\mathscr{G}'$. 显然，$\mathscr{G}'$是严格正则的. 根据 (2)，$\mathscr{G}'$中的节点个数不超过 $2n$. 由于 $\mathscr{G}$ 的正则性，$\mathscr{G}$ 中的单连隐节点不会超过一半，从而 $\mathscr{G}$ 中的节点总数 不会超过 $\mathscr{G}'$的 2 倍. 所以 $n+h\leqslant 2\times 2n$. 于是，$h\leqslant 3n$. 引理得证. □

推论 9.1 以一组给定的有穷个变量为显变量的正则预 HLC 模型的个数是有穷的.

证明 设有 n 个显变量 Y_1，Y_2，…，Y_n. 用 $N(k)$ 记以 Y_1，Y_2，…，Y_n 为显变量且含有 k 个隐变量的正则预 HLC 模型之个数. 显然，$N(k)$ 不超过由 $n+k$ 个给定节点形成的树的个数，而后者明显是有限的. 根据引理 9.3(1)，以 $Y_1,Y_2,\cdots,Y_n$ 为显变量的正则预 HLC 模型中最多有不超过 $3n$ 个隐变量. 于是，以 $Y_1,Y_2,\cdots,Y_n$ 为显变量的正则预 HLC 模型的个数不超过 $\sum_{k=1}^{3n}N(k)$，因此推论得证. □

定理 9.4 以一组给定的有穷个变量为显变量的正则 HLC 模型的个数是有穷的.

证明 这是引理 9.1 和推论 9.1 的一个明显推论. □

9.6　多层隐类模型学习算法

隐类模型中只有一个隐变量，它的每一个取值表示一个隐类，这些隐类放在一起形成了对所有样本的一个聚类. HLC 模型中有多个隐变量，每一个隐变量对应于一种聚类方式. 因此，用 HLC 模型分析数据是对样本同时进行**多维聚类**(multidimensional clustering). 作为例子，假设通过对中学生各科成绩进行分析获得了包含隐变量“分析能力”和“语言能力”的 HLC 模型. 那么，我们就按照分析能力将学生聚成了数类，同时又按照语言能力将学生聚成了另外数类. 与多维聚类的概念相对应，用隐类模型分析数据是将样本进行**单维聚类**，或者说**简单聚类**.

下面考虑利用爬山法寻找最优 HLC 模型的问题. 由定理 9.3，这个问题的搜索空间可以定为以数据所涉及的变量为显变量的所有正则 HLC 模型的集合. 根据定理 9.4，这个空间是有限的.

下面介绍**双重爬山法**（doubly hill climbing）算法，简称 **DHC 算法**. 它把模型优化问题分解成如下两个子问题：

（1）势学习问题：给定一个正则预 HLC 模型，怎样优化其隐变量的势?

（2）预模型学习问题：怎样寻找最优正则 HLC 模型所对应的预模型?

我们看到，DHC 算法不是直接在所有正则 HLC 模型的空间中进行搜索，而是利用势学习和预模型学习这两个问题把该空间分成了两个层次，并在每一个层次中分别进行搜索. 这就是它为什么称为双重爬山法的原因.

9.6.1　势学习算法

设 $\mathscr{G}^0$ 是势学习问题中给定的正则预 HLC 模型. 势学习问题的搜索空间是所有 $\mathscr{G}^0$ 所能生成的正则 HLC 模型的集合. 图 9.11 给出了一个在这个空间中利用爬山法进行搜索，从而确定隐变量的势的算法，称为 LearnCard-HC. 搜索从所有隐变量的势都是 2 的那个模型开始. 在搜索的每一步，LearnCard-HC 通过修改当前模型而得到一系列的候选模型，用于修改当前模型的搜索算子是将某一隐变量的势加 1，如果修改所得的候选模型不是正则的，则将之舍去. 然后，LearnCard-HC 计算每一个候选模型的评分，选择评分最高者作为下一步搜索的出发点. 在计算模型评分时，需要估计模型的参数. 为此，LearnCard-HC 调用 EM 算法. 搜索如此循环进行，直到通过修改当前模型不能再找到评分更高的模型为止.

LearnCard-HC($\mathbf{Y}, \mathscr{D}, \mathscr{G}^0, f, \delta$)

输入：$\mathbf{Y}$——一组显变量；　　$\mathscr{D}$——一组关于 $\mathbf{Y}$ 的数据；

$\mathscr{G}^0$——一个以 $\mathbf{Y}$ 为显变量的正则预 HLC 模型；

f——一个模型评分函数；　　δ——EM 算法收敛阈值.

输出：一个正则 HLC 模型；

```
1: 𝒢 ← 由 𝒢⁰ 生成的、每个隐变量的势均为 2 的 HLC 模型；
2: θ ← EM(𝒢, 𝒟, δ); oldScore ← f(𝒢, θ | 𝒟);
3: while (true)
4:     𝒢* ← null; θ* ← null; newScore ← −∞;
5:     for(每个通过将 𝒢 中某一隐变量的势加 1 所能得到的正则 HLC 模型 𝒢′)
6:         θ′ ← EM(𝒢′, 𝒟, δ);
7:         tempScore ← f(𝒢′, θ′ | 𝒟);
8:         if(tempScore > newScore)
9:             𝒢* ← 𝒢′; θ* ← θ′; newScore ← tempScore;
10:        end if
11:    end for
12:    if(newScore > oldScore)
13:        𝒢 ← 𝒢*; θ ← θ*; oldScore ← newScore;
14:    else
15:        return(𝒢, θ);
16:    end if
17: end while
```

图 9.11　势学习爬山算法

9.6.2　模型学习算法

预模型学习问题的搜索空间是以数据中涉及的变量为显变量的正则预 HLC 模型的集合. 图 9.12 给出了一个在这个空间中利用爬山法进行搜索而获得一个正则 HLC 模型的算法，称为 LearnHLCM-DHC. 搜索从预 LC 模型开始. 在搜索的每一步，LearnHLCM-DHC 对当前预模型进行修改，获得一系列候选正则预模型. 修改当前预模型时用到 3 个搜索算子，即**加节点**（node introduction）、**减节点**（node deletion）和**移节点**（node relocation）. 对这些算子在稍后将会有详细讨论. 对每一个候选正则预模型，LearnHLCM-DHC 调用 LearnCard-HC 来优化其中隐变量的势并估计其参数，得到一个候选模型. 然后，它计算每一个候选模型的评分并选择分数最高的作为下一步搜索的出发点. 搜索如此循环进行，直到通过修改当前模型无法再找到评分更高的模型为止. 下面逐个介绍这 3 个搜索算子.

LearnHLCM-DHC($\mathbf{Y}, \mathscr{D}, f, \delta$)

输入：$\mathbf{Y}$——一组显变量；　$\mathscr{D}$——一组关于 $\mathbf{Y}$ 的数据；
　f——一个模型评分函数；　δ——EM 算法收敛阈值；

输出：一个正则 HLC 模型：

```
1: 𝒢⁰ ← 以 Y 为显变量的预 LC 模型;
2: (𝒢, θ) ← LearnCard-HC(Y, 𝒟, 𝒢⁰, f, δ);
3: oldScore ← f(𝒢, θ | 𝒟);
4: while (true)
5:     𝒢* ← null; θ* ← null; newScore ← −∞;
6:     for(每个对 𝒢 的预模型进行一次加节点、减节点或移节点运算所能得到的正则预 HLC 模型 𝒢⁰)
7:         (𝒢′, θ′) ← LearnCard-HC(Y, 𝒟, 𝒢⁰, f, δ);
8:         tempScore ← f(𝒢′, θ′ | 𝒟);
9:         if(tempScore > newScore)
10:            𝒢* ← 𝒢′; θ* ← θ′ ; newScore ← tempScore;
11:        end if
12:    end for
13:    if(newScore > oldScore)
14:        𝒢 ← 𝒢* ; θ ← θ* ; oldScore ← newScore;
15:    else
16:        return(𝒢, θ);
17:    end if
18: end while
```

图 9.12　学习 HLC 模型的双重爬山算法

1. 加节点

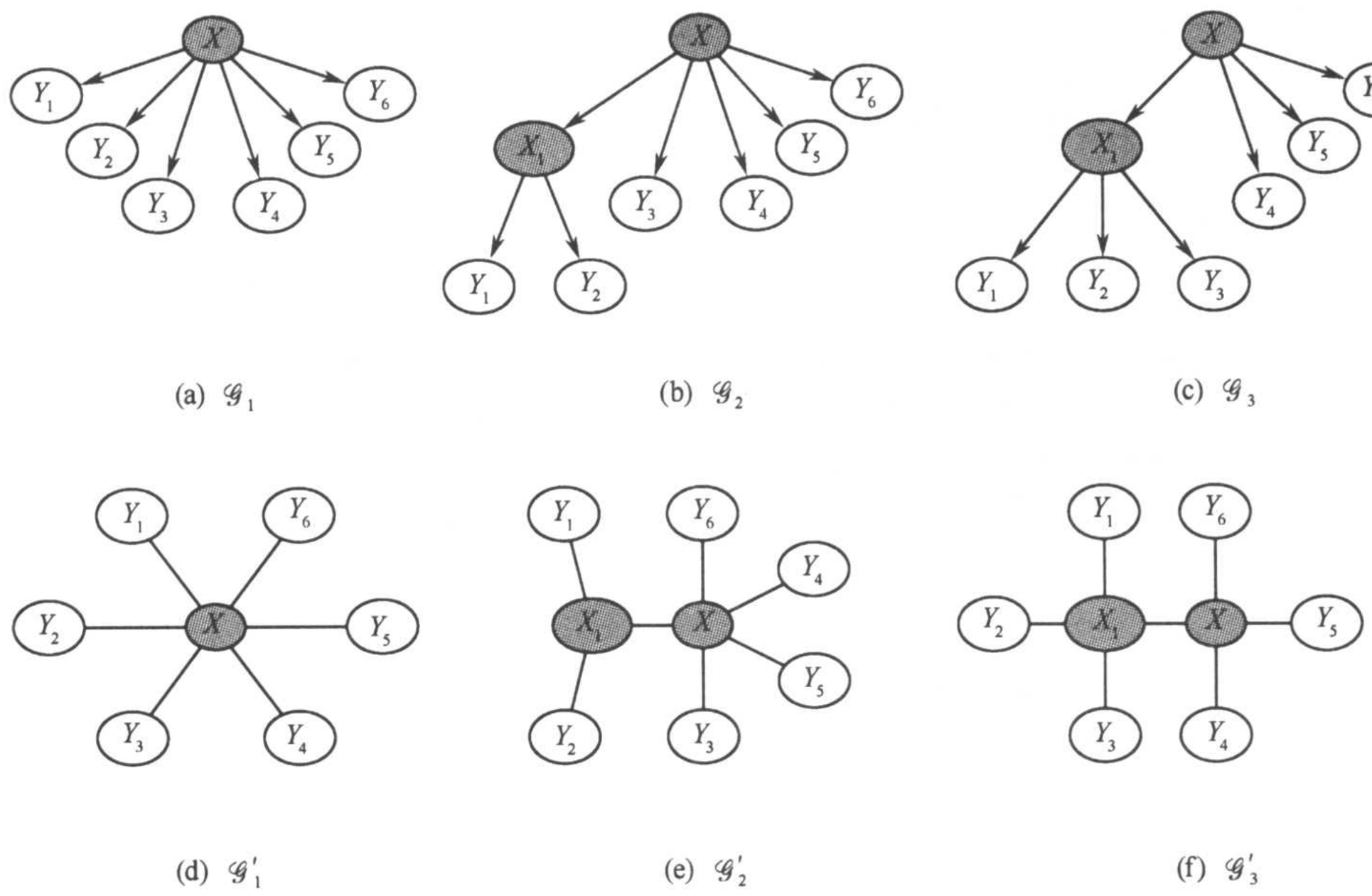
为什么要引入加节点这个算子呢？为了回答这个问题，先来看看有根模型的情况．考虑图 9.13 中的模型 $\mathscr{G}_1$．它假设，给定 X 时，所有显变量 Y_1，Y_2，…，Y_6相互条件独立．此假设不一定与实际相符．例如，给定 X 时，Y_1 和 Y_2 可能不是相互独立的．这时，为使模型更符合实际，可以对它进行修改．一个自然的修改办法就是引入一个新的隐变量 X_1，得到新模型 $\mathscr{G}_2$．在 $\mathscr{G}_2$ 中，给定 X 时，Y_1 和 Y_2 由于受 X_1 的影响，仍然相互依赖．比起 $\mathscr{G}_1$，$\mathscr{G}_2$ 与实际可能更为相符．

把上述思想引入到无根模型中，就得到加节点这个搜索算子．设 X 是一个隐节点，Z_1，Z_2 是与 X 相邻的两个节点．加入一个新的隐节点 Z 把 X 与 Z_1 和 Z_2 分开即为

(1) 去掉 X 与 Z_1 和 Z_2 之间的边；

(2) 引入一个新的隐节点 Z，并在 Z 与 X，Z_1 和 Z_2 之间分别加一条边．

在图 9.13 所示的模型 $\mathscr{G}_1'$中，加入新节点 X_1 把 X 与 Y_1 和 Y_2 分开，就得到模型 $\mathscr{G}_2'$．



(a) $\mathscr{G}_1$　(b) $\mathscr{G}_2$　(c) $\mathscr{G}_3$

(d) $\mathscr{G}'_1$　(e) $\mathscr{G}'_2$　(f) $\mathscr{G}'_3$

图 9.13　搜索算子示例

一个隐节点 X 往往有多个相邻节点. 出于对算法复杂度的考虑，LearuHL-CM-DHC 只考虑引入新节点来把 X 与它的两个相邻节点分开，而不考虑把 X 与它的任意多个相邻节点分开. 设 X 有 k 个相邻节点，这个限制使得对于 X 只有 $\frac{k(k-1)}{2}$ 种引入新节点的方法. 而如果没有 这个限制，则有 2^k 种方式. 后面将会看到，尽管加上了这个限制，加节点算子配合减节点和移节点仍能保证对搜索空间的遍历性.

对于一个正则预 HLC 模型使用加节点算子，所得到的可能不是正则预 HLC 模型. 这样的情况有两种：一种是当隐节点 X 只有两个相邻节点 Z_1 和 Z_2 时，若通过加节点将 X 与 Z_1 和 Z_2 分开，之后 X 就只有一个相邻节点，从而结果不再是预 HLC 模型，因为预 HLC 模型中隐变量不能是叶节点. 另一种情况是当 X 有 3 个相邻节点 Z_1，Z_2 和 Z_3，且其中之一，如 Z_3，是一个单连隐节点时，若引入新的隐节点将 X 与 Z_1 和 Z_2 分开，则所生成的新模型将有两个相邻的单连隐节点，即 X 和 Z_3，这破坏了预模型的正则性. 算法 LearnHLCM-DHC 禁止在上述两种情况下使用加节点算子.

2. 减节点

设 X 是一隐节点，Z_1，Z_2，…，Z_k 是与 X 相邻的节点. 如果其中之一，如 Z_1，也是隐节点，那么就可以**相对于 Z_1 消去 X**，即

(1)从模型中除去 X；

(2)在 Z_1 与 Z_2，Z_3，…，Z_k 之间分别加一条边.

在图 9.13 所示的模型 $\mathscr{G}_2'$ 中，相对于 X 消去 X_1，就还原到模型 $\mathscr{G}_1'$.

3. 移节点

设隐节点 X 的一个相邻节点 Z 也是隐节点，并设 Z' 是另外一个与 X 相邻的节点. **将 Z' 从 X 处移到 Z 处**即为

(1) 去掉 Z' 与 X 之间的边；

(2) 在 Z' 与 Z 之间加一条边.

在图 9.13 所示的模型 $\mathscr{G}_2'$ 中，将 Y_3 从 X 处移到 X_1 处，就得到 $\mathscr{G}_3'$. 而在 $\mathscr{G}_3'$ 中，将 Y_3 从 X_1 处移到 X 处就得到 $\mathscr{G}_2'$.

注意一个节点只能在两个相邻隐节点之间移动. 在图 9.10 中，不可以将 Y_1 从 X_2 处移到 X_3 处，因为 X_2 和 X_3 不相邻. 与加节点的情况类似，这个限制也是出于对计算复杂度的考虑.

对于一个正则预 HLC 模型，使用移节点算子所得到的可能不是正则预 HLC 模型. 这样的情况有两种：第一种情况是当隐节点 X 只有两个相邻节点 Z_1 和 Z_2，且 Z_2 也是隐节点时，若将 Z_1 从 X 处移到 Z_2 处，之后 X 就只有一个相邻节点，从而结果不是正则预 HLC 模型；第二种情况是当 X 有 3 个相邻节点 Z_1，Z_2 和 Z_3，其中 Z_2 和 Z_3 是隐节点，且 Z_3 是单连节点的时候，若将 Z_1 从 X 处移到 Z_2 处会造成两个相邻的单连隐节点，即 X 和 Z_3. 这破坏了预模型的正则性. LearnHLCM-DHC 算法禁止在上述两种情况下使用移节点算子.

移节点算子的效果可以用加节点和减节点的组合来实现. 在图 9.13 中，通过加、减节点，可以 从 $\mathscr{G}_2'$ 得到 $\mathscr{G}_3'$：首先加入隐变量 X_2 将 X 与 X_1 和 Y_3 分开，然后相对于 X_2 将 X_1 消去，就得到 $\mathscr{G}_3'$. 既然移节点算子等价于加节点和减节点的一个组合，那么为什么还要将它作为第三个算子引入呢？这样做的主要原因是为了减少 LearnHLCM-DHC 算法停在局部最优的机会.

下面证明上述 3 个算子对搜索空间的遍历性.

定理 9.5　考虑以一组给定变量 **Y** 为显变量的所有正则预 HLC 模型所组成的集合 **S**. 从 **S** 中任何一个模型出发，利用加节点、减节点以及移节点算子，可以到达 **S** 中其它任何一个模型，并且不经过 **S** 以外的模型.

证明　以 **Y** 为显变量的正则预 LC 模型在 **S** 中，记之为 $\mathscr{G}_0$. 为了证明定理，

只需 证明如下命题：从 $\mathscr{G}_0$ 出发利用加节点、减节点以及移节点算子；可以到达 **S** 中任何一个模型 $\mathscr{G}$，且不经过 **S** 以外的模型；反之亦然.

对 $\mathscr{G}$ 中的隐变量的个数进行归纳. 当 $\mathscr{G}$ 只包含一个隐变量时，$\mathscr{G}$ 就是 $\mathscr{G}_0$，命题在平凡意义上成立. 假设命题在 $\mathscr{G}$ 包含 h 个隐变量时成立，下面考虑 $\mathscr{G}$ 包含 $h+1$ 个隐变量的情况. 这时，$\mathscr{G}$ 中一定有两个相邻的隐节点 X_1 和 X_2，而且它们之中至少有一个不是单连节点. 不失一般性，设 X_1 不是单连节点. 那么，除 X_2 以外，X_1 至少还有两个相邻节点. 设 Z_1，Z_2，Z_3，…，Z_k 是 X_1 的除 X_2 以外的相邻节点. 考虑先将 Z_3，…，Z_k 从 X_1 处移到 X_2，然后相对于 X_2 消去 X_1. 记这样得到的模型为 $\mathscr{G}'$. 由于 $\mathscr{G}'$ 只包含 h 个隐变量，根据归纳假设，$\mathscr{G}_0$ 和 $\mathscr{G}'$ 可以不经 **S** 以外的模型而互达. 而根据 $\mathscr{G}'$ 的构造过程，$\mathscr{G}'$ 和 $\mathscr{G}$ 也可以不经 **S** 以外的模型而互达. 所以命题在 $h+1$ 个隐变量时也成立. 定理得证. □

9.6.3 复杂度分析

下面分析 LearnHLCM-DHC 算法在最坏情况下的时间复杂度. 设有 n 个显变量. 根据引理 9.3 (1)，正则预 HLC 模型中的隐变量的个数小于 $3n$，从而变量总数小于 $4n$.

在搜索的每一步，LearnHLCM-DHC 对当前模型进行加节点、减节点或移节点，从而生成一系列候选预模型. 加节点算子涉及某个隐变量的两个相邻节点，因此它所生成的候选预模型的个数不会超过 $\frac{4n(4n-1)}{2}<8n^2$. 减节点算子涉及某个待除去的隐节点和它的一个相邻隐节点，因此它所生成的预模型的个数不会超过 $3n\cdot(3n-1)<9n^2$. 移节点算子涉及某个隐节点的两相邻节点，且其中之一必为隐节点，所以，它生成的预模型的个数不会超过 $4n\cdot 3n=12n^2$. 因此，在搜索的每一步，LearnHLCM-DHC 所构造的候选预模型的个数不会超过 $29n^2$. 设搜索步数不超过 N. 那么，LearnHLCM-DHC 分析的预模型总数不 超过 $29n^2N$.

对于每一个预 HLC 模型，LearnHLCM-DHC 调用 LearnCard-HC 来确定其中隐变量的势. LearnCard-HC 也是一个爬山算法. 它首先考虑所有隐变量的势都是 2 的模型，然后开始搜索. 在搜索的每一步，它通过把当前模型中的某个隐变量的势加 1 的方式，生成一系列候选模型. 因为隐变量的总数不超过 $3n$，所以 LearnCard-HC 在每一步最多只生成 $3n$ 个候选模型. 设 k 是隐变量的势的一个上界. 那么 LearnCard-HC 的搜索步数不超过 $3nk$，从而它所构造的模型的个数不超过 $3n\cdot 3nk=9n^2k$.

对于每一个候选模型，LearnCard-HC 调用 EM 算法来估计它的参数. 在候选模型中进行一次概率推理的时间 是 $O(4n)$. 设有 m 份不同的数据样本. 那么，

EM 的每一次迭代花的时间是 $O(4nm)$. 设 EM 算法的迭代次数不超过 M. 那么，它用的总时间不超过 $O(4nmM)$.

综上所述，LearnHLCM-DHC 算法在最坏情况下的时间复杂度为

$$O(29n^2N \cdot 9n^2k \cdot 4nmM) = O(1044MNkmn^5). \tag{9.24}$$

9.6.4 仿真试验

为了评估 LearnHLCM-DHC 算法，Zhang（2004）做了一些仿真实验. 实验所使用的原生模型如图 9.7 所示，其中所有变量的势均设为 3，而模型参数是随机产生的. 实验首先从这个模型中随机抽样，得到 4 组训练数据，其样本量分别是 5000、10000、50000 和 100000. 同时还抽得 5000 个测试样本（testing data）. 随机抽样得到的每一个样本都包含所有显变量的状态.

Zhang（2004）利用 LearnHLCM-DHC 分别对上述 4 组训练数据进行了分析，所用的评分函数有 4 种，即 BIC 评分、AIC 评分、（原始）CS 评分和 HVL 评分. 在描述 LearnHLCM-DHC 算法时，图 9.13 假设使用的评分函数是罚项似然度. 由于 CS 评分和 HVL 评分不是罚项似然度，在使用它们时，算法需要有一些相应的改动. 在使用 HVL 评分时，Zhang（2004）从训练数据中分出 25% 作为验证数据.

实验中，EM 算法的设置如下：在模型优化过程中，EM 的收敛阈值设为 0.01，而最后在估计最优模型的参数时，其收敛阈值设为 0.0001. EM 的迭代次数的上限设为 200. 为了避免陷入局部最优，实验使用（Chickering and Heckerman，1997）所描述的多初始点方法：从 R 个随机产生的初始参数值出发，进行 1 次 EM 步骤，然后淘汰 $R/2$ 个似然度低的参数值；对剩下的 $R/2$ 个参数值，进行 2 次 EM 步骤，然后淘汰 $R/4$ 个似然度低的参数值；对剩下的 $R/4$ 个参数值，进行 4 次 EM 步骤，然后淘汰 $R/8$ 个似然度低的参数值；如此下去直至只剩一个参数值，然后对它不断进行 EM 步骤，直至收敛或 EM 步骤数目达到设定的上限. 在 Zhang（2004）的实验中，R 设为 64.

LearnHLCM-DHC 算法是用 Java 实现的，实验在一台 Pentium III 1GHz 的机器上进行. 就时间而言，使用 BIC 评分函数分析样本量为 10000 的训练数据花了 97 个小时. 算法在其它情况下的运行时间都处于同一个数量级.

仿真实验的目的是要看看 LearnHLCM-DHC 通过分析数据所得到的模型与数据的原生模型是否基本相同. 这包括两方面：一是它们所表示的关于显变量的联合概率分布是否基本相同？二是它们的结构及隐变量的势是否一致？

第一个问题可以通过比较基于测试数据的对数似然度来回答①. 图 9.14 给出了 LearnHLCM-DHC 算法所获得的模型以及数据原生模型基于测试数据的对数似然度. 我们看到，使用不同评分函数所获得的结果的对数似然度差别甚微，而且都接近原生模型的对数似然度. 这说明，这些结果模型所表示的关于显变量的概率分布与原生模型所表示的分布基本相同.

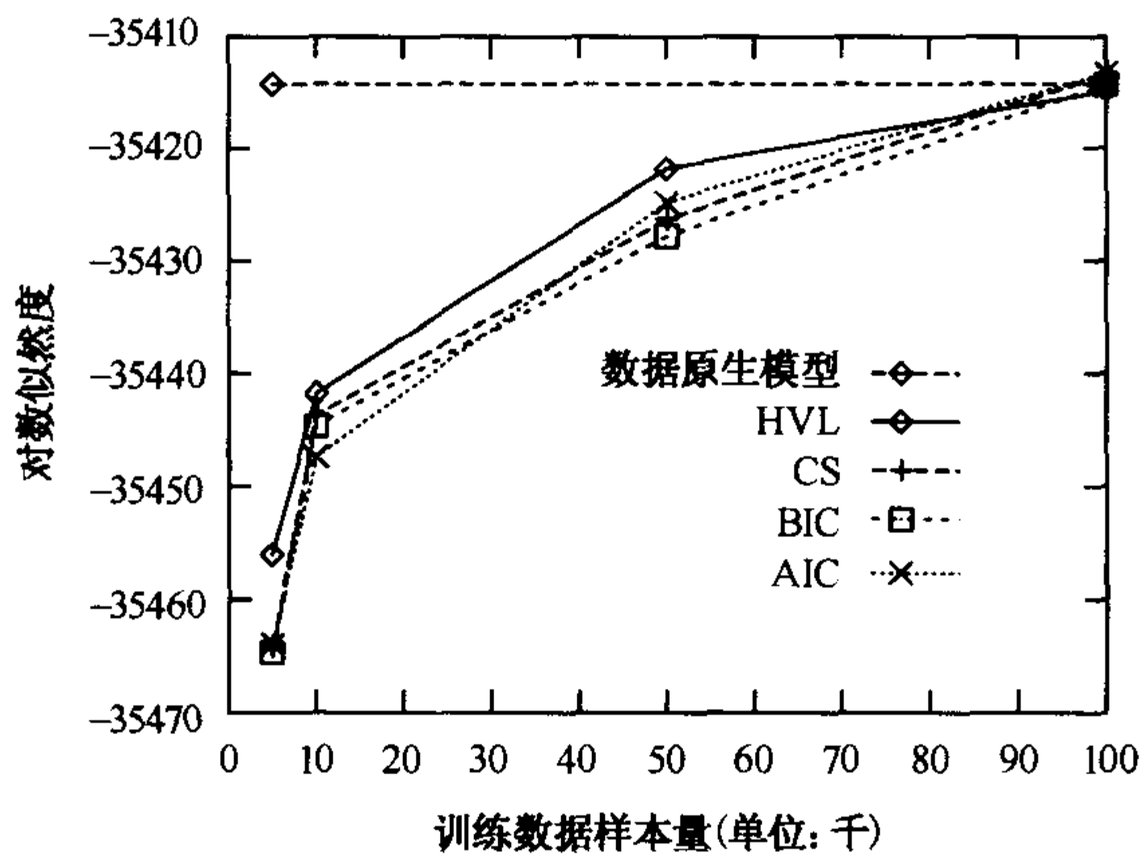


图 9.14 结果模型及数据原生模型基于测试数据的对数似然度

接下来看看模型结构. 表 9.2 给出了 LearnHLCM-DHC 算法在不同情况下获得的模型的结构，以及从这些结构出发到达原生结构所需使用的搜索步数. 其中 $\mathscr{G}_0$ 是原生模型之结构，$\mathscr{G}_1$，$\mathscr{G}_2$，…，$\mathscr{G}_6$ 由图 9.15 给出. LearnHLCM-DHC 算法的结果是无根模型. 图 9.15 把模型的结构画成有根树，这主要是为了阅读上的方便.

表 9.2 结果模型的结构以及它们与原生模型结构之间的距离

	5×10^3	10×10^3	50×10^3	100×10^3
BIC	1 ($\mathscr{G}_1$)	1 ($\mathscr{G}_1$)	0 ($\mathscr{G}_0$)	0 ($\mathscr{G}_0$)
CS	1 ($\mathscr{G}_1$)	2 ($\mathscr{G}_2$)	0 ($\mathscr{G}_0$)	0 ($\mathscr{G}_0$)
HVL	2 ($\mathscr{G}_3$)	4 ($\mathscr{G}_6$)	0 ($\mathscr{G}_0$)	0 ($\mathscr{G}_0$)
AIC	3 ($\mathscr{G}_4$)	4 ($\mathscr{G}_5$)	0 ($\mathscr{G}_0$)	0 ($\mathscr{G}_0$)

① 用 $\mathbf{Y}$ 记所有显变量的集合. 用 $P(\mathbf{Y})$ 记 $\mathbf{Y}$ 在原生模型中的分布，$P_{\mathcal{N}}(\mathbf{Y})$ 记它们在学习所得模型中的分布. 用 $\mathscr{D}_t$ 记测试数据，其样本量为 m. 设 $\hat{P}(\mathbf{Y})$ 是 $\mathbf{Y}$ 基于 $\mathscr{D}_t$ 的经验分布. 那么有 $\log P(\mathscr{D}_t)=m\sum_{\mathbf{Y}}\hat{P}(\mathbf{Y})\log P(\mathbf{Y})$，$\log P_{\mathcal{N}}(\mathscr{D}_t)=m\sum_{\mathbf{Y}}\hat{P}(\mathbf{Y})\log P_{\mathcal{N}}(\mathbf{Y})$. 所以，$\log P(\mathscr{D}_t)-\log P_{\mathcal{N}}(\mathscr{D}_t)=m\sum_{\mathbf{Y}}\hat{P}(\mathbf{Y})\log\frac{P(\mathbf{Y})}{P_{\mathcal{N}}(\mathbf{Y})}$. 当 m 趋于无穷时，$\hat{P}$ 几乎必然收敛于 P. 于是，$\log P(\mathscr{D}_t)-\log P_{\mathcal{N}}(\mathscr{D}_t)$ 几乎必然收敛于 $mKL(P, P_{\mathcal{N}})$，也就是说，$\log P(\mathscr{D}_t)-\log P_{\mathcal{N}}(\mathscr{D}_t)$ 量度的是 $P(\mathbf{Y})$ 与 $P_{\mathcal{N}}(\mathbf{Y})$ 间的 KL 距离.

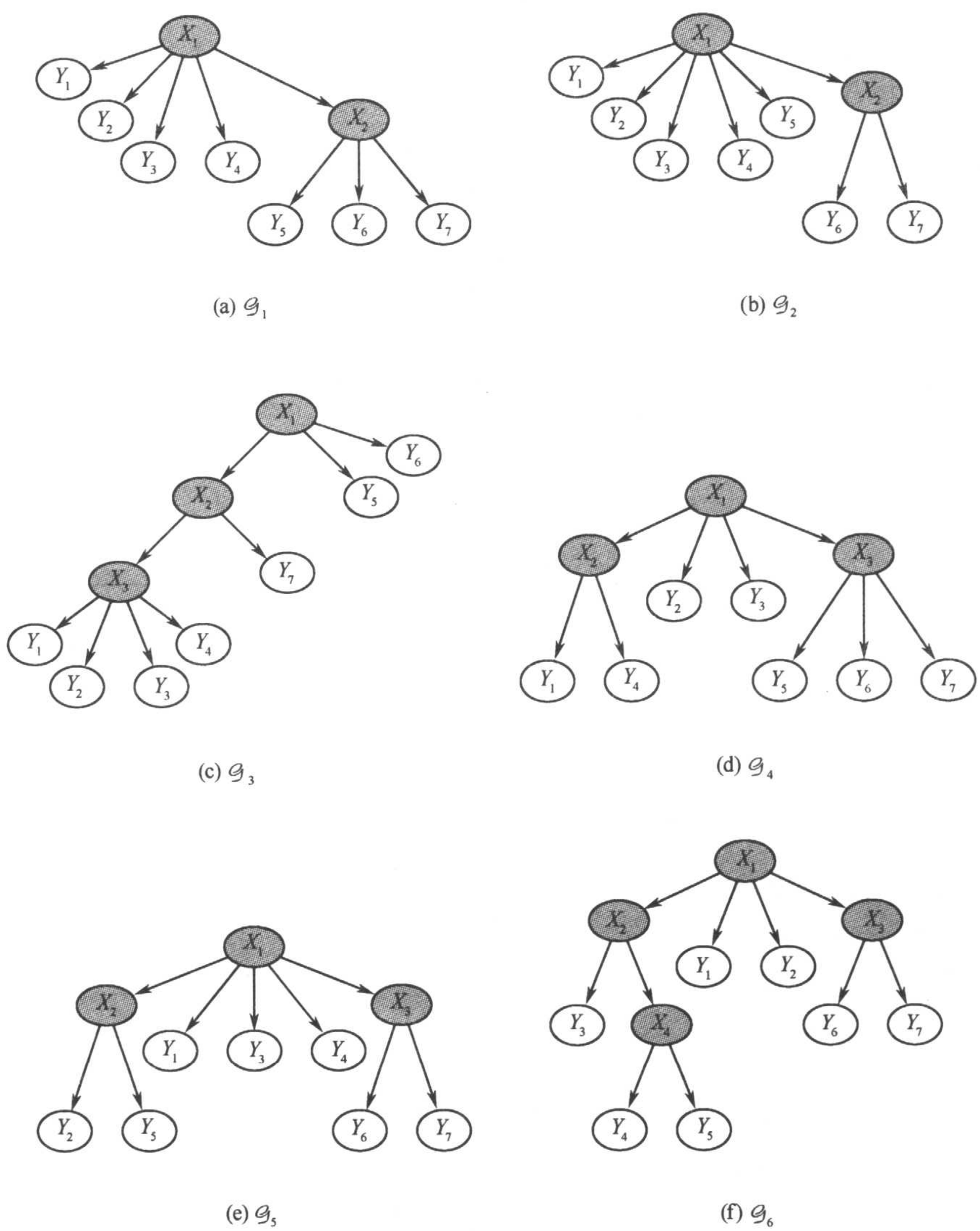


图 9.15　一些结果模型的结构

我们看到，当使用 BIC 或 CS 评分函数时，LearnHLCM-DHC 从样本量为 50000 和 100000 的两组数据中获得了原生模型的结构. 在样本量为 5000 和 10000 时，它获得的模型结构与原生模型结构 $\mathcal{G}_0$ 接近. 特别是在使用 BIC 评分时，LearnHLCM-DHC 得到 $\mathcal{G}_1$，它与 $\mathcal{G}_0$ 只差一步：在 $\mathcal{G}_1$ 中引入一个隐变量将 X_1 与 Y_4 和 X_2 分开就得到 $\mathcal{G}_0$. 因此，LearnHLCM-DHC 将 $\mathcal{G}_0$ 和 $\mathcal{G}_1$ 进行过直接

的比较. 它选择了 $\mathscr{G}_1$ 而非 $\mathscr{G}_0$ 是因为在样本量为 5000 和 10000 时，$\mathscr{G}_1$ 的评分比 $\mathscr{G}_0$ 更高.

当使用 AIC 和 HVL 评分时，LearnHLCM-DHC 也从样本量为 50000 和 100000 的两组数据中获得了原生模型的结构 $\mathscr{G}_0$. 但是，在样本量为 5000 和 10000 时，它获得的模型结构就与原生模型结构有明显差别.

尽管在样本量为 50000 和 100000 时，无论使用哪种评分函数，LearnHLCM-DHC 都找到了原生模型的结构，但是，结果模型中隐变量的势却因评分函数而异. 如表 9.3 所示，BIC 和 CS 给出的隐变量的势的估计更接近真实值，且当样本量增大时有所改善，而 AIC 和 HVL 给出的某些估计大于真实值，且当样本量增加时，估计精度并无改善.

表 9.3 一些结果模型中隐变量的势

	50000				100000			
	BIC	CS	HVL	AIC	BIC	CS	HVL	AIC
X_1	2	2	2	2	2	2	2	2
X_2	3	3	3	3	3	3	3	4
X_3	2	2	4	4	3	3	4	4

在上述的实验中，原生模型的参数是随机生成的. 这样，有的变量与其父节点之间的关联可能很弱. 因此，Zhang（2004）做了第二个实验，在原生模型中在每对相邻变量之间的关联都比较强的情况下测试 LearnHLCM-DHC. 在这个实验中，模型参数仍是随机生成的，只是要求每个条件分布至少有一个概率值大于 0.6. 其它设置与第一个实验完全一样.

就基于测试数据的对数似然度而言，第二个实验的结果与第一个实验大体一致. 但在模型结构及隐变量的势方面却有明显分别. 当使用 BIC 和 CS 评分函数时，LearnHLCM-DHC 从所有 4 组数据中都获得了原生模型的结构，变量 X_2 和 X_3 的势也总能估计正确. 变量 X_1 的势在样本量为 100000 的情况下得到了正确的估计，而在其它情况下估计为 2. 这些都远远优于第一个 实验的结果.

另一方面，当使用 AIC 和 HVL 评分时，即便在样本量为 50000 和 100000 的原生模型的结构中，LearnHLCM-DHC 的表现却不如第一个实验.

9.7 文献介绍

隐类分析是由 Lazarsfeld 和 Henry（1968）首先提出的，最新发展可参见有关文献（Hagenaars and Mccutcheon，2002）. Bartholomew 和 Knott（1999）论

述了隐类分析与因子分析（factor analysis）、隐特质分析（latent trait analysis）以及隐概况分析（latent profile analysis）之间的关系．在隐类分析中，显变量和隐变量都是离散的．在隐特质分析中，显变量是离散的而隐变量是连续的，它主要应用于教育测验和心理测试．在隐概况分析中，显变量是连续的而隐变量是离散的．在机器学习中，它被称为混合模型（mixture models）（McLachlan and Basford，1988）．在因子分析中，显变量和隐变量都是连续的，它与主成分分析（principal component analysis）（Jolliffe，1988）关系密切．

在贝叶斯网文献方面，Elidan et al.（2000）和 Elidan and Friedman（2001）讨论了如何在离散变量贝叶斯网中引入隐变量并决定其势．Spirtes et al.（2001）和 Silva et al.（2003）在结构方程模型（structured equation models）（Bollen，1989）的框架下研究了利用连续变量贝叶斯网揭示隐结构的问题．在上述文献中也有隐变量模型与隐结构模型的概念，但它们与本书所定义的概念有所不同．

本章关于 HLC 模型的内容取材于 Zhang（2004）．HLC 模型与生物信息学中的进化树（phylogenetic tree）（Durbin et al.，1988）关系密切．事实上，进化树的概率模型可以视为 HLC 模型的一种特殊情况．Zhang 等（2004）将 HLC 模型的基本思想应用到分类问题中，提出了多层朴素贝叶斯模型（hierarchical naïve Bayes model）．Zhang 和 Kocka（2004a）证明了一个有助于计算 HLC 模型有效维数的结果．Chen 等（2005）把这个结果推广到了包含隐变量的多亲树（polytree）模型．LearnHLCM-DHC 算法的复杂度很高，只能分析包含 10 个左右显变量的数据．最近算法研究有新进展，已经可以处理 100 个左右显变量，详情参见作者的网页。

第四部分

贝叶斯网应用

第 10 章　隐结构模型与中医辨证

第 9 章讨论了怎样通过数据分析获得隐结构模型. 隐结构模型在诸如心理学、社会学等领域有众多应用. 本章介绍它在中医研究中的一个应用例子，内容取材于文献（张连文、袁世宏，2004).

辨证论治是中医的精髓. 然而，辨证却缺乏客观定量标准，在临床实践中受到主观因素的严重影响. 本章考虑怎样为中医辨证建立客观定量标准的问题，并介绍一种旨在解决这个问题的研究方法. 由于它的主要特点是使用隐结构模型，故称其为**隐结构法**.

10.1　中医辨证的客观化、定量化

中医诊断是首先通过望、闻、问、切收集病情资料，然后对病情资料进行综合分析，进而对疾病的病因、病位、病势、病性等本质做出判断. 判断的结果称为**证**或**证候**，而判断的过程称为**辨证**.

就现状而言，中医辨证有 3 个特点. 首先，中医辨证没有客观标准；其次，医师辨证技能的高低很大程度取决于经验，只有高龄医师才会受到充分信任；第三，辨证结果受主观因素的影响很大，不同医师对同一病人所做的辨证结论往往不同. 因此，十分有必要为中医辨证建立客观的、定量的标准.

10.1.1　相关工作回顾

关于中医辨证客观化、定量化的研究已经有相当长的历史，学者们做了大量工作. 下面先简要回顾这些工作.

1. 证实质研究

证实质研究[1]兴起于 20 世纪 50 年代，“这一研究的近期目标是，寻找和确定中医证的客观检测指标，对证进行定量表达；远期目标是，用客观检测指标对疾病做出完全的证的诊断”（梁茂新等，1998). 证实质研究是否成功关键要看是否能够找到既具有敏感性又具有特异性的检测指标. 敏感性是指当病人出现某证

---

① 这里，证实质研究主要指寻找理化特异指标的辨证微观化研究，不包括动物模型研究.

候时，指标应该呈阳性；特异性是指病人无该证候时，指标应该呈阴性. 这里病人是否有某证候由专家根据中医辨证理论和经验来判断.

如果证实质研究的目标实现了，辨证自然就有了客观标准. 遗憾的是，证实质研究并不成功. 以肾虚证研究为例：不同的学者从不同的组织系统和生理病理角度选择了上百个指标，但目前还没有发现一个相对肾虚具有特异性的指标（王洪琦等，1999（第一章，第五节）). “24 小时尿 17-羟皮质类固醇降低”曾一度被认为是肾阳虚的特异性指标，但后来发现，它在脾阳虚时也会出现. 对其它证候的研究也有相同的情况. 有鉴于此，赵国求（1999）说：“40 年来，中医证实质研究已经将成百上千的实验数据摆在我们面前，按照原定研究目标，这些不争气的数据给予我们的只是困惑和迷茫.”专家们对证实质研究未能成功的原因进行了深刻分析，主要有：

（1）证实质研究所遵循的思想和方法实际上是西医的方法，它和中医是两种不同的学术体系，两者在认识论和方法论上截然不同，不属于同一个认识层次（杨维益等，1996）；

（2）证是一种涉及全身多系统的功能病理状态，其概念欠清晰，难以定性、定量、定位（董昌武，廖圣宝，高尔鑫，2000). 因此目前多数学者认为，证候不能用类似西医的方法靠几个特异性的指标来判定.

2. 回归分析和判别分析在证研究中的应用

有众多学者把回归和判别分析应用于中医证的计量诊断研究，目的是在症状与证候之间建立起统计定量关系. 这种研究的出发点是：一组已由专家分类的病例，每份病例不但有病人的症状，而且还有专家对病人所做的辨证结论，学者们利用标准统计方法对病例数据进行分析，获得回归方程或判别函数，其中，症状为自变量，而证候为因变量. 如刘和朱（1998）对中医产科脾气虚证 57 例和非脾气虚证 23 例进行了分析，建立了如下回归方程：

$$Y = 11.324 + 7.632X_{33} + 4.593X_{30} + 1.597X_{61} + \cdots + 0.567X_{26}.$$

这里所选入的自变量分别是滑胎（X_{33}）、胎漏（X_{30}）、脉象虚弱（X_{61}）、…、妊娠腹痛（X_{26}），它们取值为 1 或 0，分别代表症状的出现或不出现. 当因变量 $Y \geqslant 37.54$ 时判断为脾气虚证，而当 $Y < 37.54$ 时判断为非脾气虚证. 又如李小兵等（2000）从已由专家分类的数据出发为心脑血管病痰证建立了如下判别函数：

$$Y_1 = -8.9596 + 2.5393X_1 + 3.6459X_2 + \cdots + 3.9959X_9,$$

$$Y_2 = -1.5726 + 1.1467X_1 + 2.0832X_2 + \cdots + 1.1244X_9.$$

这里所选入的自变量分别是肥胖（X_1）、胸腹痞闷（X_2）、…、脉滑（X_9），其取值也是 1 或 0. 若 $Y_1 > Y_2$，则判为痰证；若 $Y_1 < Y_2$，则判为非痰证.

如上所述的回归方程和判别函数可以用作辨证的定量标准. 但是，由于它们是对数据的数学总结，而每份数据中因变量（证候）的取值是专家的主观判断，所以它们其实是专家行为的数学总结. 因此，用它们作为辨证标准其实就是用专家作为标准，缺乏客观性.

3. 聚类分析在证研究中的应用

辨证就是把病人进行分类，同一证候的病人分到同一类，不同证候的病人分到不同的类，分类的根据是病人的症状. 那么，这些类是否客观存在呢？为了回答这个问题，可以首先进行流行病学调查收集病例，按症状对病人进行聚类，症状相近的聚为同一类，症状不相似的聚到不同的类；然后把所得的类与证候进行比较. 如果一个类 C 中的病人的症状与某证候 S 的症状整体上吻合，那么类 C 就可以看成是对有证候 S 的病人的群体的抽样. 这不但证实了证候 S 的客观存在，而且为建立辨证标准提供了基础：对于一个新来的病人，可以将其症状与类 C 中的病人的症状进行比较，如果相似则断定病人有证候 S，若否则断定病人无证候 S.

基于上述思想，杨学鹏（1993）（第九章，第五节）对八纲辨证中的阴阳两证进行了研究. 遗憾的是结果不甚理想，聚类分析所得到的类与阴证、阳证的特性并不吻合. 就这次研究本身而言，失败的原因可能很多. 比方所用的是基于距离的聚类分析，也许距离定义得不合理，也许距离聚类这一方法本身不适宜. 一般来讲，简单使用聚类分析来研究证有难于克服的实质性困难. 这主要是因为辨证是一个**多维同时分类问题**. 例如，按不同气虚程度可以把病人分为几类，同时，按不同血虚程度又可以把病人分为另外几类，等等. 这就好比可以同时把人按性别分为“男”、“女”，按年龄分为“老”、“中”、“青”，按职业分为“工人”、“农民”等. 因为简单聚类方法是单维的，所以它无法处理多维问题.

10.1.2　隐结构法

上面的讨论显示，中医辨证的客观定量标准难以通过证实质研究来建立；回归分析和判别分析可以建立定量的辨证标准，但缺乏客观性；简单聚类方法是单维的，无法处理辨证的多维性. 本章探讨利用隐结构模型为辨证建立客观定量标准的可能性.

1. 辨证理论与隐结构

中医辨证有其理论基础，那就是中医辨证理论. 既然如此，为什么辨证没有客观标准呢？这是因为中医辨证理论还没有得到证实，而且缺乏系统性和定量

性. 上面讲到的证实质研究和聚类分析的目的就是要为辨证理论提供证实并将之定量化，但是没有成功. 其实，这是一个十分困难的问题. 要了解它之所以困难的根本原因，首先需要对辨证理论的内容有个基本的认识.

中医辨证理论讲述的是证候与症状之间的关系. 例如关于肾阴虚，中医认为：

> “肾阴亏虚，脑髓、官窍、骨骼失养，则见腰膝酸痛、眩晕耳鸣、健忘、齿松发落；……；虚火上扰心神，故烦热少寐；肾阴不足，失于滋润，虚火蕴蒸，故见口燥咽干，形体消瘦，潮热盗汗，或骨蒸发热，颧红，尿黄少. 舌红少苔，脉细数，为阴虚内热之象.”（朱文锋，1995）

这里，诸如“腰膝酸痛”、“舌红少苔”、“脉细数”的症状是可以通过望闻问切而直接观察到的，是显变量. 而诸如“肾阴虚”的证候迄今没有办法对它们进行直接观察，是隐变量. 除了证候以外，诸如“肾阴虚失养”、“肾阴虚失滋润”、“阴虚内热”的病理因子也是隐变量. 隐变量与隐变量之间以及隐变量与显变量之间的关系构成了一个隐结构.

中医辨证理论所描述的是一个把各式各样的症状联系起来的隐结构. 隐结构是不可能直接观测到的，这是中医辨证理论难以科学证实和科学定量化的根本原因.

2. 隐结构法的基本思想

隐变量是什么？这是一个哲学问题，有几种不同的答案（见 10.3 节）. 本章采用如下观点（Sobel，1994）：隐变量是人脑思维的产物；人脑发明隐变量、隐结构的目的是要解释在众多事例中所观察到的规律. 比方在中学生人群中，人们观察到：一个数学成绩好的学生，物理、化学成绩往往也好；反之亦然. 人们于是引入“分析能力”这个隐变量来解释这一现象：分析能力强的学生，数学、物理、化学 3 门课程的成绩一般都较好；而分析能力弱的学生，这 3 门成绩往往都较差. 在中医临床实践中，“舌上少津”、“口渴喜饮”、“小便涩少”这 3 个症状常常结伴出现. 中医理论引入“津亏”这个隐变量来解释这一现象：津亏能导致这 3 个症状的出现；而无津亏时，它们则一般不会出现.

基于上述观点可以认为，中医理论是人脑在通过对众多病例的观察发现了一些规律后，为解释这些规律而引入的一个用自然语言描述的隐结构. 中医临床用人脑构造的隐结构来指导辨证，因此缺乏客观性、定量性.

本章介绍一种旨在为辨证建立客观、定量标准的研究方法，即隐结构法. 它的基本思想是用电脑来取代人脑进行数据分析构造隐结构. 具体地说就是：首先进行流行病学调查，系统收集病例；然后用计算机对所得数据进行分析，找

出规律，并且构造隐结构模型来解释这些规律；最后用所得的隐结构模型来指导辨证.

用电脑构造的隐结构来指导辨证的好处是辨证客观性可以得到大大提高，而且定量化也不再是一个问题. 除了客观性和定量性以外，更重要的是辨证质量. 后面读者将会看到，用电脑构造的隐结构来指导辨证，辨证的质量也可以得到保证和提高.

隐结构法实质上是一种聚类分析方法. 在上面的例子中，引入“分析能力”这个隐变量意味着按“分析能力”这个隐指标把学生聚成数类，每一类对应“分析能力”的一个“程度”①. 同样，引入“津亏”这个隐变量意味着按“津亏”这个隐指标把病人聚成数类，每一类对应“津亏”的一个“程度”. 用隐结构模型分析数据会引入多个隐变量，每一个隐变量都是一隐聚类指标，对应一种聚类方式. 所以，隐结构法是一种多维同时聚类方法.

3. 隐结构法与肾虚辨证

作为隐结构法的一个示例，张连文、袁世宏（2004）用 HLC 模型对肾虚辨证进行了研究. 本章主要介绍这项研究的情况：10.2 节介绍数据收集情况. 为方便有中医背景的读者，10.3 节直观地解释用 HLC 模型分析数据的原理. 10.4 节介绍数据分析情况. 第 10.5 和 10.6 两节把数据分析所得的 HLC 模型 M^* 与相关中医理论分别在定性和定量两个层次进行比较，发现两者基本吻合，说明 M^* 是一个高质量模型. 第 10.7 节论述怎样用模型 M^* 指导辨证及其意义. 在模型 M^* 指导下的辨证称为**模型辨证**. 第 10.8 节通过案例分析揭示模型辨证与专家（组）辨证的异同，发现双方在实质上是基本一致的. 第 10.5、10.6、10.8 三节的比较表明，模型辨证的质量是可以信赖的. 因为模型辨证具有客观性和定量性，而且它的质量又可以信赖，所以它具备了成为客观定量辨证标准的条件.

10.2　肾虚数据收集

在进行流行病学调查之前，需要确定调查涉及哪些症状. 依据《国家证候诊疗标准》(GB/T16751.2)(1997) 和中医诊断学教材（朱文锋，1995），张连文、袁世宏（2004）选用了如表 10.1 所示的 67 个肾虚症状变量，它们反映了肾脏藏精主水纳气、主生殖、主骨生髓、司二便、开窍于耳、其华在发、齿为骨之余等各个方面的生理功能和病理变化. 表中的变量名都以“程度”两字结

① 这里，“程度”二字加上了引号是因为聚类分析得到的类并不是总可以诠释为程度.

尾. 比方第一个变量其实是“精神不振程度”，“程度”二字被省略是为了排版方便.

表 10.1 调查涉及的症状变量

1. 精神不振	2. 腰膝酸困	3. 下肢困乏不便	4. 足跟困痛	5. 头部昏晕
6. 视力减退老花	7. 听力减退	8. 耳鸣	9. 失眠	10. 记忆力减退
11. 淡漠寡言	12. 计算能力减退	13. 头发白或脱	14. 鼻毛阴毛白	15. 齿松脱落
16. 心慌心烦	17. 咳嗽喘息	18. 声低气怯	19. 畏寒怕冷	20. 头部怕冷
21. 四肢发冷	22. 腰背发冷	23. 阴部冷感	24. 自汗	25. 面目浮肿
26. 下肢浮肿	27. 盗汗	28. 五心烦热	29. 身潮热	30. 颧部赤热
31. 久患它病史	32. 面色萎黄	33. 面色淡白	34. 面色灰暗	35. 腰痛喜按
36. 痛处拒按	37. 有刺痛感	38. 口渴喜饮	39. 咽干舌燥	40. 夜尿频
41. 昼尿频	42. 小便余沥不尽	43. 小便涩少	44. 昼小便失禁	45. 夜遗尿
46. 尿细小清长	47. 尿黄赤短少	48. 便秘	49. 便稀	50. 完谷不化
51. 五更泻	52. 大便失禁	53. 舌上少津	54. 苔白滑润	55. 舌质红
56. 舌质紫暗	57. 舌质淡	58. 苔黄	59. 苔薄	60. 苔厚
61. 脉细数	62. 脉迟	63. 脉沉	64. 脉浮	65. 尺脉不足
66. 脉有力	67. 脉无力			

变量的程度分为“无”、“轻”、“中”、“重”四级. 为保证可操作性、一致性，减少人为误差，调查过程采用了严石林、王米渠等（2001）所制定的肾虚症状轻重程度标准. 这个标准的建立考虑到了症状表象轻重程度、伴随条件、持续时间、发作频率和诱发条件 5 个因素，是经过参照对比、专家咨询、集体讨论、小样本测试以及最后修正等一系列工作而完成的. 该标准关于精神不振程度、腰膝酸困程度和夜间尿频程度的规定如下：

精神不振程度：“无”——精力旺盛，基本无疲乏感觉；“轻”——劳累后出现精神不振，休息即可恢复；“中”——无原因地出现精神不振，休息后也不容易恢复；“重”——成天无精打采，做事提不起精神，疲惫不堪，休息后也不能恢复.

腰部酸困程度：“无”——没有任何酸困疼痛不适感；“轻”——偶尔发生，遇劳后加重，可以久坐；“中”——经常感到腰困乏力，双腿无力，不耐久坐久立；“重”——腰膝困乏酸痛，不能坐立行走.

夜间尿频程度：“无”——0～1 次；“轻”——2～3 次；“中”——4～5 次；“重”——6 次以上.

抽样调查还需要划定一个样本空间. 有研究显示，在自然人群中虚证患病率

仅为 7.66%（陆金宝等，2002）．如果以自然人群为样本空间，那么相对于揭示肾虚证的规律这一目标，多数样本将只是起陪衬作用，因而需要的样本量将会很大．由于资源上的限制，决定暂时先考虑 60 岁或以上的老年人这一肾虚证常常出现的人群．注意，这个选择意味着研究结论只适用于这个子人群，而不一定适用于整个自然人群．

表 10.2 列出了组织调查的主要人员．其他参加调查的人员包括博士、硕士研究生和部分学习成绩优秀的本科生．调查前，对调查人员进行了培训．培训内容包括中医基础理论中有关肾脏的理论，症状程度判别标准，流行病学调查的原则、方法和注意事项等．调查在老年人集中的群落场所进行，如敬老院、老年大学、老年活动中心、社区保健点、离退休职工处，甚至老年人较集中的农村院落等．如果场所有管理者，就直接和管理者进行联系并由管理者组织和协助．如场所没有明确管理者，就选择有影响和有能力的热心老年人协助召集组织．在调查过程中，每个调查人员依次逐个对老年人进行调查询问．如果有本科学生参加时则必配备有指导老师现场指导，及时解决现场临时出现的一些具体问题，对学生把握不准的评判进行指导．如果现场解决仍存有疑问，留待该次调查结束后统一讨论确定．每一份调查表均要求填写调查者的姓名，并于结束调查后统一检查登记核实．

表 10.2　调查所得原始资料的基本情况

地点	人群	主持人	例数
成都	（敬老院）	袁世宏	316
太原	（老年大学、敬老院）	袁世宏	167
山西	（县、乡农村）	袁世宏	1117
重庆	（退休工人）	吴斌	300
成都	（福利院）	吴斌	122
内蒙	（工厂、农村）	朱永光	578
		总计：	2600

在调查过程中为保证资料的准确可靠，排除了不能清晰进行问答的老人．另外，由于社会文化的原因，有关性生殖方面的资料获取较难而且不准确，况且大部分老年人都已经没有这方面的生活，因此这部分资料一般未获取．在最后的资料筛选中，如果发现症状缺失较多，或者书写质量较差的病例则删去不用．调查总计获得数据 2600 例．

10.3 数据分析原理

第 9 章给出了一个学习 HLC 模型的算法，同时指出了还有其它算法. 对于中医研究人员来说，算法细节并不重要，重要的是直观原理. 下面先解释 HLC 模型学习的直观原理，然后讨论隐变量的语义.

用 HLC 模型来分析数据即是假设显变量的背后是一个 HLC 模型，并且基于数据对这个模型进行推测. 数据中只有显变量的取值，而无隐变量的取值. 换句话说，没有对隐结构的直接观测. 既然如此，为什么可以对隐结构进行推测呢？这是因为在数据中能够找到隐结构的踪迹. 例如，在前面关于中学生的例子中，隐变量“分析能力”同时影响数学、物理、化学 3 门成绩. 在数据中，这种影响表现为数理化这 3 门成绩的相互关联：数学成绩好则其它两门成绩往往也好，反之亦然.

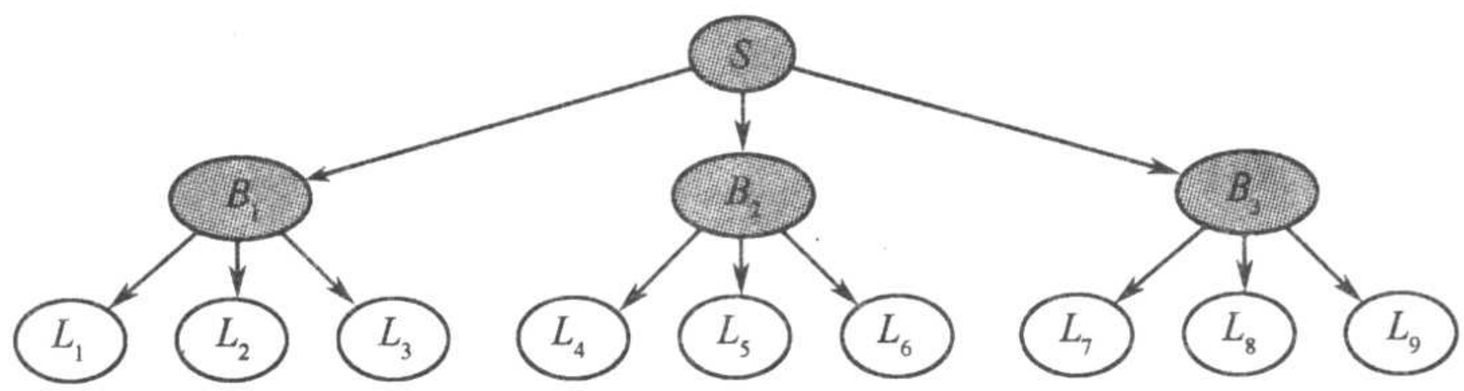
既然隐结构在数据中有迹可寻，那么就可以通过分析数据来对它们进行推测. 下面用一个例子来具体说明这一点.

例 10.1（彩灯原理） 想象 9 盏由一个电子控制器所控制的彩灯，把它们分别记为 $L_1, L_2, \cdots, L_9$. 每盏灯或“亮”或“灭”. 假设对这些彩灯的亮灭进行了长时间观察，现在希望对它们背后的控制器的逻辑结构做一些推测.

观察发现 L_1，L_2，L_3 要么一齐亮，要么一齐灭. 基于这一点，可以推测这 3 盏灯属于同一电路支线，记之为 B_1：当 B_1“通电”时，L_1，L_2，L_3 都亮；而当 B_1“不通电”时，L_1，L_2，L_3 都灭. 观察还发现 $L_4 \sim L_6$ 和 $L_7 \sim L_9$ 这两组灯也分别同时亮或灭，于是推测它们分别属于另外两条支线. 用 B_2，B_3 分别表示这两条支线，它们和 $L_4 \sim L_6$，$L_7 \sim L_9$ 之间的关系与 B_1 和 $L_1 \sim L_3$ 之间的关系相同.

观察还发现，1/3 的时间 $L_1 \sim L_6$ 同时亮，另外 1/3 的时间 $L_4 \sim L_9$ 同时亮，其它 1/3 的时间所有灯都灭. 于是推测控制器有 3 个不同状态，“关”、“开 1”和“开 2”，且它每个状态呆 1/3 时间. 当处于“关”状态时，它不给任何支线送电；当处于“开 1”状态时，它给支线 B_1，B_2 送电，但不给支线 B_3 送电；当处于“开 2”状态时，它给支线 B_2，B_3 送电，但不给支线 B_1 送电.

上面的推测结果是一个 HLC 模型. 该模型有 4 个隐变量：控制器的状态 S，以及 3 条支线的通电情况 B_1，B_2，B_3. S 有 3 个可能取值“关”、“开 1”和“开 2”；其它隐变量有 2 个可能取值“通电”和“不通电”. 模型的结构如图 10.1 所示. 图 10.1 同时也给出了模型的几个概率分布. 建议读者自己写出模型的其余概率分布.



$P(S)$

S=关	S=开 1	S=开 2
1/3	1/3	1/3

$P(B_1 \mid S)$

	B_1=不通电	B_1=通电
S=关	1	0
S=开 1	0	1
S=开 2	1	0

$P(B_2 \mid S)$

	B_2=不通电	B_2=不通电
S=关	1	0
S=开 1	0	1
S=开 2	0	1

$P(L_1 \mid B_1)$

	L_1=灭	L_1=亮
B_1=不通电	1	0
B_1=通电	0	1

$P(L_2 \mid B_1)$

	L_2=灭	L_2=亮
B_1=不通电	1	0
B_1=通电	0	1

$P(L_3 \mid B_1)$

	L_3=灭	L_3=亮
B_1=不通电	1	0
B_1=通电	0	1

图 10.1　基于彩灯亮与灭的规律推测彩灯控制器的结构

通过推测得到的模型是不是控制器结构的真实写照？这不得而知．但是，它很好地解释了观察到的规律，而且是具有这一特性的最简单模型．因此，从统计学角度看，它是一个高质量的模型．□

症状在某病人身上出现与否就好比彩灯在某一时刻的亮与灭．通过对彩灯的长时间观察，发现不同彩灯的亮与灭不是相互独立的，而是相互联系的，这使得可以对彩灯控制器的结构进行推测．同样，通过对症状在众多病人身上出现情况的观察，可以发现症状之间各式各样的统计关联关系，这些关系为寻找症状背后的隐结构提供了基础．当然此时问题远比例 10.1 复杂，所以需要求助于第 9 章介绍的理论和算法．

用隐结构分析数据首先要假设显变量的背后有一个隐结构模型．在这个假设的模型中，隐变量的语义（semantics）是什么？它们是客观存在、主观构造？还是其它？关于这个问题，学者们进行了不少讨论，有许多不同的观点（Borsboom et al.，2003；Sobel，1994）．最主要的有两种观点，即**实体观**（realism）

和**构造观**（constructivism）. 实体观认为隐变量是像年龄、性别那样客观存在的实体（real entity），而构造观则认为隐变量是人为构建的、是为方便把握复杂现象特点而引入的主观概念. 在实体观下，隐变量之所以是隐变量有两种不同的原因：或者是由于它未被观测（unobserved），或者是由于它不可观测（unobservable）. 学习隐结构的过程是寻找真理的过程，学习而得的模型可以谈对错：若它与客观存在的隐结构一样，则是正确的；否则，是错误的. 在构造观下，隐变量之所以是隐变量是因为对它无所谓观测. 对于学习而得的模型也无正确或错误可言，因为没有判断对错的标准. 人们关心的只是学到的模型从数据角度看是否恰当（empirical adequacy），所以这可以说是一种实用主义的观点. 本章采用的是构造观.

10.4 肾虚数据分析

张连文、袁世宏（2004）利用 Zhang 和 Kocka（2004b）提出的启发式单重爬山法对第 10.2 节所描述的肾虚数据进行了分析，得到一个 HLC 模型 M^*. 本节介绍数据分析的情况.

在技术层面讲，分析数据的目的是要寻找最优 HLC 模型，即 BIC 分数最高的 HLC 模型. 为什么在分析肾虚数据时要把视野限制在 HLC 模型的范围内，而不考虑更为复杂的隐结构模型呢？这有两个原因：首先，相对于现存的 LC 模型，HLC 模型已经进行了很大推广，而且正如后面将会看到的那样，HLC 模型已经可以揭示数据中所隐含的许多规律. 其次，HLC 模型处理起来已经相当困难，关于它的许多问题还亟待研究，在这种情况下去考虑更为复杂的模型是不适宜的.

怎样寻找最优 HLC 模型呢？张连文和袁世宏（2004）首先使用 Zhang and Kocka（2004b）的启发式单重爬山法进行自动搜索. 这个算法暂时无法处理数据所涉及全部 67 个变量，于是他们选择了其中对肾虚辨证最重要的 35 个来进行分析. 分析过程在一台 2.4GHz 的奔腾 4 计算机上用了 98.5 小时完成. 得到的模型记为 M_0^*，它的 BIC 分是 -73947.

爬山法有一个共同的缺点：它们有可能停于局部最优. 所以，M_0^* 未必是 BIC 分数最高的 HLC 模型. 为了使结果的质量尽可能地高，张连文和袁世宏从 M_0^* 出发进行了人工搜索：首先利用背景知识对 M_0^* 进行多方面修改，从而得到一系列新模型；然后计算出每一个新模型结构的 BIC 分，并且把它们与 M_0^* 的 BIC 分进行比较，选择分数最高的模型作为最后结果. 最后所得到的模型记为 M^*，它的 BIC 分数是 -73877，而它的结构如图 10.2 所示. M^* 的 BIC 分数略高于 M_0^*，它的结构也与 M_0^* 相近.

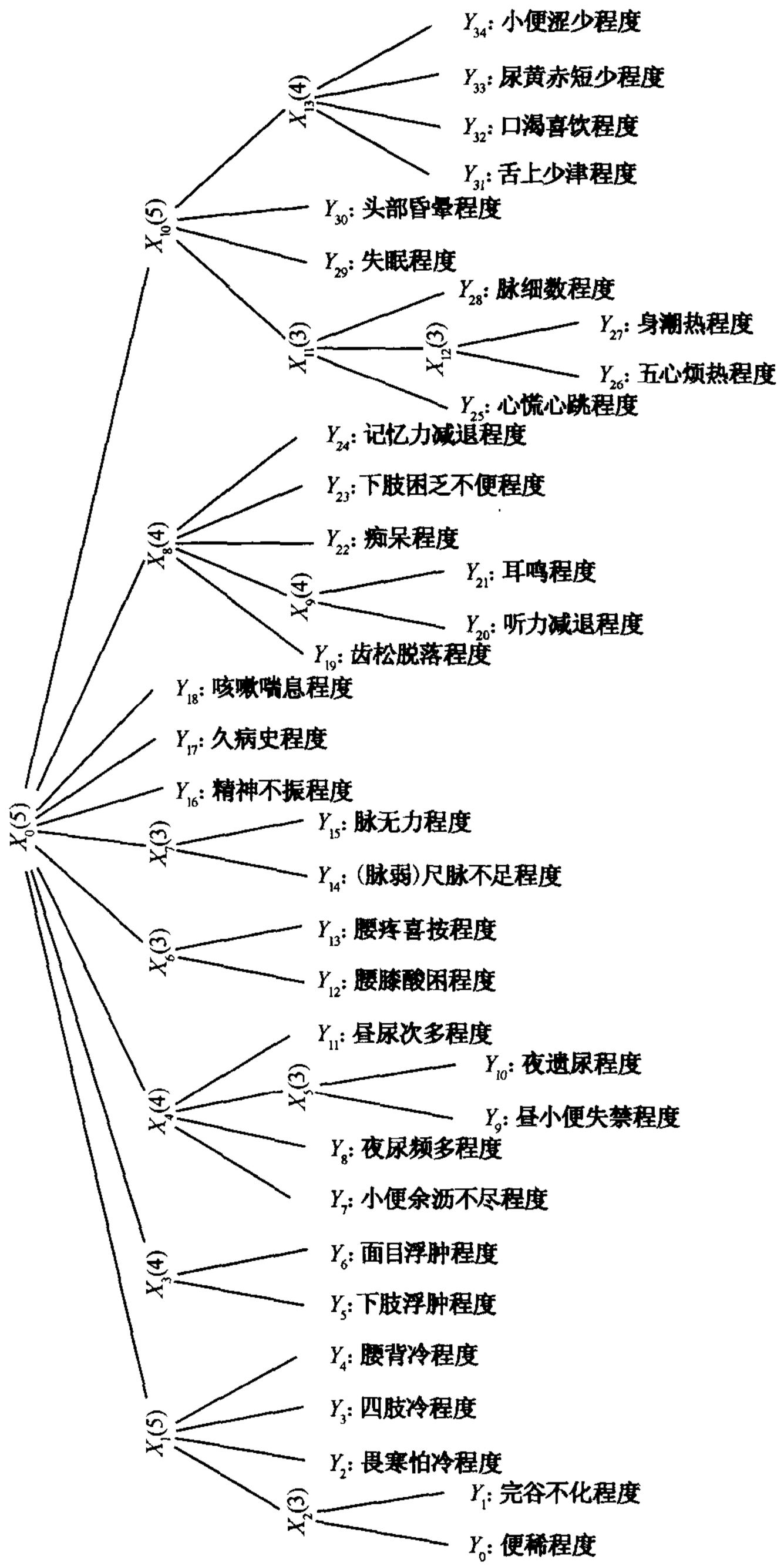

Y_{34}: 小便涩少程度
Y_{33}: 尿黄赤短少程度
Y_{32}: 口渴喜饮程度
Y_{31}: 舌上少津程度
$X_{13}(4)$
Y_{30}: 头部昏晕程度
Y_{29}: 失眠程度
$X_{10}(5)$
Y_{28}: 脉细数程度
Y_{27}: 身潮热程度
Y_{26}: 五心烦热程度
$X_{12}(3)$
$X_{11}(3)$
Y_{25}: 心慌心跳程度
Y_{24}: 记忆力减退程度
Y_{23}: 下肢困乏不便程度
Y_{22}: 痴呆程度
$X_8(4)$
Y_{21}: 耳鸣程度
$X_9(4)$
Y_{20}: 听力减退程度
Y_{19}: 齿松脱落程度
Y_{18}: 咳嗽喘息程度
Y_{17}: 久病史程度
Y_{16}: 精神不振程度
$X_0(5)$
Y_{15}: 脉无力程度
$X_7(3)$
Y_{14}: (脉弱)尺脉不足程度
Y_{13}: 腰疼喜按程度
$X_6(3)$
Y_{12}: 腰膝酸困程度
Y_{11}: 昼尿次多程度
Y_{10}: 夜遗尿程度
$X_5(3)$
Y_9: 昼小便失禁程度
$X_4(4)$
Y_8: 夜尿频多程度
Y_7: 小便余沥不尽程度
Y_6: 面目浮肿程度
$X_3(4)$
Y_5: 下肢浮肿程度
Y_4: 腰背冷程度
Y_3: 四肢冷程度
$X_1(5)$
Y_2: 畏寒怕冷程度
Y_1: 完谷不化程度
$X_2(3)$
Y_0: 便稀程度

图 10.2　模型 M^* 的结构

模型 M^* 是否最优？这一点尚有待证明．不过，张连文和袁世宏（2004）从 M^* 出发，做了进一步的自动和人工搜索，没有发现更好的模型．另外后面将会看到，M^* 已经是关于肾虚辨证的一个高质量模型．

隐结构法将来可能被用来分析各种不同的数据．人们自然要问：从两组关于同一领域的数据出发，会不会得到相同的模型？要回答这个问题，必须考虑数据背后的样本空间．假设有一个足够大的样本空间，想象从中不断抽样，因而获得越来越多的数据．在这个过程的开始，最优模型一般会随着数据量的增加而改变．但是，当数据量达到一定程度后，最优模型会稳定下来，不再改变．所以，如果两组不同的数据来自同一样本空间，而且数据量足够大，那么它们对应的最优模型就是一样的；否则，就可能不同．

上一段讨论的是最优模型相对于数据的稳定性．与之可以相提并论的是最优模型相对于变量的稳定性：给定一组数据，假设先考虑其中部分变量，相应有一个最优模型；然后增加所考虑的变量，相应有另外一个最优模型．这两个模型之间存在什么关系，它们是否在某种意义下相容（consistent）？这是一个有趣的问题．不过，迄今还没有关于它的研究．

10.5 结果模型定性内容的质量

模型 M^* 是关于肾虚辨证这个领域的一个统计学模型．它的质量怎么样？一般地说，一个模型质量的高低可以从 5 个方面来看：

（1）模型是否贴切地反映了实际情况（accuracy）；

（2）模型的内容是否清晰（clarity）；

（3）模型是否完备（completeness）；

（4）模型是否简洁（conciseness）；

（5）模型是否有自相矛盾的地方（consistency）．作为统计学模型，M^* 是清晰的，完备的，简洁的，不自相矛盾的．为了判断 M^* 是否贴切地反映了肾虚辨证的实际情况，本节和下一节将分别在定性和定量两个层次把它与相关中医理论进行比较．比较结果发现两者基本吻合，说明 M^* 是一个高质量的模型．

1. 模型 M^* 与中医理论在定性层面的比较

本小节在定性层面把模型 M^* 与肾虚辨证理论进行比较．我们从模型的左下角开始：有一个隐变量 X_1，它直接影响畏寒怕冷、四肢冷、腰背冷这 3 个症状的出现和轻重程度，同时又通过隐变量 X_2 间接影响便稀、完谷不化的出现和轻重程度．这与中医理论是否吻合？隐变量 X_1 和 X_2 的意义是什么？中医理论认为，肾藏元阳内寓命火，既可温煦脏腑经络、推动激发脏腑功能活动，又能温肢

体暖周身. 当肾阳不足失温煦时则可见畏寒肢冷、腰背怕冷. 同时肾阳虚衰火不温脾土，则可见完谷不化，大便溏薄而稀. 由此可见，在这个局部 M^* 与中医理论完全吻合，它们讲的完全是同一件事，即：有一个隐因子，它直接与畏寒怕冷、四肢冷、腰背冷这 3 个症状有关；同时它还通过另一隐因子间接地与便稀、完谷不化有关. 中医理论为这两个隐因子命了名，分别叫肾阳虚失温煦和火不温脾土. 所以，隐变量 X_1 可解释为“肾阳虚失温煦程度”，而 X_2 可解释为“火不温脾土程度”.

在 X_1 的右边，M^* 指出有隐变量 X_3，它影响下肢浮肿、面目浮肿这两个症状的出现及其轻重程度. 中医理论认为，肾的阳气亏虚，气化失司，水液失于蒸化而内停泛滥，溢于上则面目浮肿，趋于下而见下肢浮肿. 所以在这一局部 M^* 与中医理论基本吻合，X_3 可解释为“肾阳虚水泛程度”.

在 X_4 及其下这个局部，M^* 指出有隐变量 X_4，它直接影响小便余沥、夜尿频多、昼尿次多这 3 个症状的出现和轻重程度，同时又通过隐变量 X_5 间接影响昼小便失禁和夜遗尿的出现及轻重程度. 中医理论认为，膀胱为洲渎之官，它的气化约束作用控制着小便开合排泄. 当膀胱气化失司，约束失职关门不利，则可见小便余沥不尽、夜尿频多、昼尿次多. 当气化约束失职较重，关门不能，则可使尿失禁固而见夜遗尿、昼小便失禁. 因此就这一局部而言，M^* 与中医理论基本吻合，其中 X_4 可解释为“膀胱失约程度”，而 X_5 可解释为“尿失禁固程度”.

注意，在中医理论中有一个与膀胱失约密切相关的证候叫做肾气不固. 它指的是肾气亏虚，封藏固摄功能失职，致使膀胱失约、精关不固、冲任失约、带脉失固而表现出的证候. 与精关不固、冲任失约、带脉失固相关的症状调查起来有一定社会文化难度，未能获得可靠数据. 因此，M^* 中没有相应的显变量，从而不难理解为什么 M^* 没有与肾气不固相对应的隐变量.

在 X_4 的右边，M^* 指出有隐变量 X_6，它影响腰膝酸困、腰痛喜按这两个症状的出现和轻重程度；而隐变量 X_7 影响尺脉不足、脉无力这两个症状的出现和轻重程度. 中医认为腰为肾之府，肾虚则腰失所养，见腰膝酸困、腰痛喜按，故 X_6 可解释为“腰膝失养程度”. 元气根源于肾，当肾气虚衰不足，无力鼓动脉气，则脉气虚弱两尺尤甚，故 X_7 可解释为“脉气无力程度”. 因此就这一局部而言，M^* 与中医学理论完全吻合.

在 X_8 及其下这个局部，M^* 指出有隐变量 X_8，它直接影响齿松脱落、痴呆、下肢困乏不便、记忆力减退这 4 个症状的出现和轻重程度，同时还通过隐变量 X_9 间接地影响听力减退、耳鸣的出现及轻重程度. 中医认为肾藏精生髓养骨充脑，开窍于耳，其华在发，齿为骨之余. 当肾精不足，则齿易摇发易脱；精髓不足，脑髓不充，则记忆力减退，甚至痴呆；髓亏骨失充养，萎软无力，则下肢行走不便，困乏无力；耳窍失养则耳鸣、听力减退. 所以在这一局部，M^* 与中

医理论基本吻合，X_8 可解释为“肾精亏虚程度”，而 X_9 则可解释为“耳窍失养程度”.

把视线移到模型的右下角. 这里 M^* 断言有一个隐变量 X_{13}，它直接影响舌上少津、口渴喜饮、尿黄赤短少、小便涩少这 4 个症状的出现及轻重程度. 根据中医理论，阴液亏虚，津液不足，上不能滋润口舌，下不能气化生尿，故可导致舌上少津、口渴、小便短赤涩少. 由此可见，在这一局部 M^* 与中医理论完全吻合，隐变量 X_{13} 可解释为“津亏程度”.

在 X_{13} 的左边，M^* 指出有隐变量 X_{11}，它直接影响心慌心跳、脉细数的出现及轻重程度，并通过隐变量 X_{12} 间接影响五心烦热、身潮热的出现及轻重程度. 中医认为，当阴液亏虚时，阴虚不能制约阳，阳必相对亢盛，阳亢则内生火热，虚热外蒸，则见五心烦热、身潮热；虚火上扰，则见心慌心跳；阴液亏虚不充脉道则脉细，血受热迫而脉数. 所以在这一局部 M^* 与中医理论基本吻合，X_{11} 可解释为“阳亢程度”，而 X_{12} 可以解释为“虚热程度”.

在 X_{11} 和 X_{13} 的上方，模型 M^* 指出有隐变量 X_{10}，它不仅与阳亢程度（X_{11}）、阴津亏程度（X_{13}）这两个隐变量有直接联系，而且还影响失眠、头晕这两个症状的出现及轻重程度. 中医认为，当阴液亏虚时，一方面滋润濡养功能必然减退，另一方面阴不制阳，阳热之气相对偏旺而生内热. 阴液亏虚，津液不足，则有 X_{13}；虚热内生，则有 X_{11}；而阴虚濡养功能减退，则脑窍失养而见失眠头晕. 所以在这一局部，M^* 与中医理论也是基 本吻合的，而 X_{10} 可以解释为“（肾）阴虚程度”.

以上讨论过的隐变量中有 7 个与隐变量 X_0 直接相连，它们是 X_1，X_3，X_4，X_6，X_7，X_8，X_{10}. 通过分析，发现这些变量分别对应肾虚的不同侧面：肾阳虚失温煦（X_1）、肾阳虚水泛（X_3）、膀胱失约（X_4）、腰膝失养（X_6）、脉气无力（X_7）、肾精不足（X_8）和肾阴亏虚（X_{10}）. 因此，隐变量 X_0 应该解释为“肾虚程度”. 从与 X_0 直接相连的 3 个症状变量的角度看，这个诠释也是合理的：中医认为，肾虚多有久病史（Y_{17}）；肾虚则元气不足精神不振（Y_{16}）；肾虚不能纳气归元，气逆于肺则咳嗽喘息（Y_{18}）.

综上所述，模型 M^* 在定性层面与中医理论基本吻合，其中的隐变量与证候的对应关系如表 10.3 所示.

表 10.3　模型 M^* 中的隐变量与中医证候的对应关系

X_0：肾虚程度	X_1：肾阳虚失温煦程度	X_2：火不温脾土程度	X_3：肾阳虚水泛程度
X_4：膀胱失约程度	X_5：尿失禁固程度	X_6：腰膝失养程度	X_7：脉气无力程度
X_8：肾精亏虚程度	X_9：耳窍失养程度	X_{10}：（肾）阴虚程度	X_{11}：阴虚阳亢程度
X_{12}：阴虚虚热程度	X_{13}：阴虚津亏程度		

2. 模型 M^* 的不足

上面的分析显示，M^* 与中医理论在方方面面都基本吻合，说明它是一个符合肾虚辨证实际的模型. 但是，它也有与中医理论不一致的地方. 首先，在中医理论中，阴虚和精亏都可以导致失眠和头晕，然而在 M^* 中失眠和头晕却只与 X_{10}（阴虚）相连. 其次，中医认为舌上少津既可能是阴虚津亏的结果，也可能是阳虚不能蒸腾气化，水津不能上润所致，然而在 M^* 中它却只与 X_{13}（阴虚津亏）相连. 第三，根据中医理论，脉细数、尿黄赤短少应该是津亏和虚热共同的结果，但在 M^* 中它们却分别只与 X_{11}（虚热）、X_{13}（津亏）相连接. 这些差异源自 HLC 模型的一个限制，那就是一个显变量只能与一个隐变量相连. 另外，肾虚的其它证型如肾气不固、肾不纳气等证候没有在 M^* 中反映出来. 前面已经提到，M^* 中没有与肾气不固相对应的隐变量，是因为没有把有关生殖性生活的症状纳入分析范围. 同样，肾不纳气没有在 M^* 中反映出来是由于除了咳嗽喘息以外，肾不纳气的其它症状，比方呼吸表浅、呼多吸少、咳嗽无力等，被排除在分析范围以外.

3. 有向模型与无向模型

细心的读者也许已经注意到，图 10.1 中连接不同节点的边是有向的，然而图 10.2 中的边却是无向的. 这是因为仅仅基于数据无法决定 HLC 模型的根，从而无法决定出模型中边的方向. 图 10.2 所示的是一个无根 HLC 模型，它的边是无向的. 另一方面，中医理论中有些关系是有向的，有因果意味. 比方津亏会导致口渴喜饮、尿黄赤短少等. 因此，把 M^* 与中医理论进行对比，可以决定出 M^* 中的一些边的方向. 举例说，既然 X_{13} 对应津亏，而且津亏导致口渴喜饮 (Y_{32})，那么 X_{13} 与 Y_{32} 之间边的方向应该是从 X_{13} 到 Y_{32}. 其实，在图 10.2 中，之所以选择 X_0 为根，完全是这种对应分析的结果. 图中除边 X_0-Y_{17} 以外，其它与症状变量直接相连的边都可以看成是有向的，方向从上到下.

10.6　结果模型定量内容的质量

本节在定量层面把模型 M^* 与中医理论进行比较. 为此，首先需要了解模型 M^* 和中医理论有什么样的定量内容.

1. 模型 M^* 的定量内容

模型 M^* 有两方面的定量内容：一方面，它明确规定隐变量的取值个数；另

一方面，它用条件概率分布来定量刻画变量之间的关系. 以隐变量 X_2 为例. 10.5 节基于它直接影响便稀、完谷不化的出现和程度这一事实，把它诠释为火不温脾土程度. 在定量层面，X_2 共有 3 个不同的取值，分别记为 s_0，s_1 和 s_2，而 X_2 与便稀程度（Y_0）和完谷不化程度（Y_1）之间的定量关系由条件概率分布 $P(Y_0 \mid X_2)$ 和 $P(Y_1 \mid X_2)$ 给出，见图 10.3. 另外，X_2 与 X_1 之间的定量关系由条件概率分布 $P(X_2 \mid X_1)$ 所刻画；X_1 与 X_0 之间的定量关系由条件概率分布 $P(X_1 \mid X_0)$ 所刻画，等等.

$P(Y_0 \mid X_2)$

Y_0	Y_0=无	Y_0=轻	Y_0=中	Y_0=重
$X_2=s_0$	0.947	0.044	0.009	0.000
$X_2=s_1$	0.422	0.569	0.005	0.004
$X_2=s_2$	0.304	0.381	0.293	0.022

$P(Y_1 \mid X_2)$

Y_1	Y_1=无	Y_1=轻	Y_1=中	Y_1=重
$X_2=s_0$	0.970	0.028	0.002	0.000
$X_2=s_1$	0.696	0.304	0.000	0.000
$X_2=s_2$	0.309	0.462	0.210	0.019

图 10.3 条件概率分布 $P(Y_0 \mid X_2)$，$P(Y_1 \mid X_2)$

为了与中医理论比较，下面从另外一个角度看模型 M^* 的定量内容. 用 HLC 模型分析数据就是对数据进行多维聚类，模型 M^* 中的每一个隐变量都是一个隐聚类指标，对应一种聚类方式. 以下把“隐变量”和“隐聚类指标”当同义词使用. 隐变量 X_2 的意义是火不温脾土程度，它有 3 个不同的取值 s_0，s_1 和 s_2. 这就是说按照火不温脾土程度这个隐指标，模型 M^* 把数据样本，或者说数据样本对应的人群，聚成了 3 类. 由于这些类不是直接观察到的，因此是隐类. 这些隐类的特性由它们的**类概率分布**（class probability distribution）所刻画. 例如，要了解 $X_2=s_2$ 这个类的特性，就要看便稀和完谷不化这两个症状在其中的分布情况，即要看类概率分布 $P(Y_0 \mid X_2=s_2)$ 和 $P(Y_1 \mid X_2=s_2)$. 如图 10.3 中的两个表的最后一行所示，$P(Y_0 \mid X_2=s_2)$ 指的是：在 $X_2=s_2$ 这个类中，有轻度、中度、重度便稀的人分别占 38%、29%、2%；$P(Y_1 \mid X_2=s_2)$ 指的是：在 $X_2=s_2$ 这个类中，有轻度、中度、重度完谷不化的人分别占 46%、21%、2%. 在图 10.3 中还能找到 $X_2=s_0$ 和 $X_2=s_1$ 两个类的类概率分布.

为了让隐类的特性一目了然，可以把类概率分布用一个直方图展示出来. 图 10.4中的第三幅图给出了 $X_2=s_2$ 这个类的类概率直方图. 图中的两个柱体分别表示在类 $X_2=s_2$ 中，症状变量 Y_0 和 Y_1 的概率分布. Y_0 和 Y_1 有 4 个取值：无、轻、中、重；与之对应，柱体分为不同灰度的 4 段，每段对应一个取值，其长度表示取该值的概率. 图 10.4 同时还给出了 $X_2=s_0$ 和 $X_2=s_1$ 两类的概率直方图.

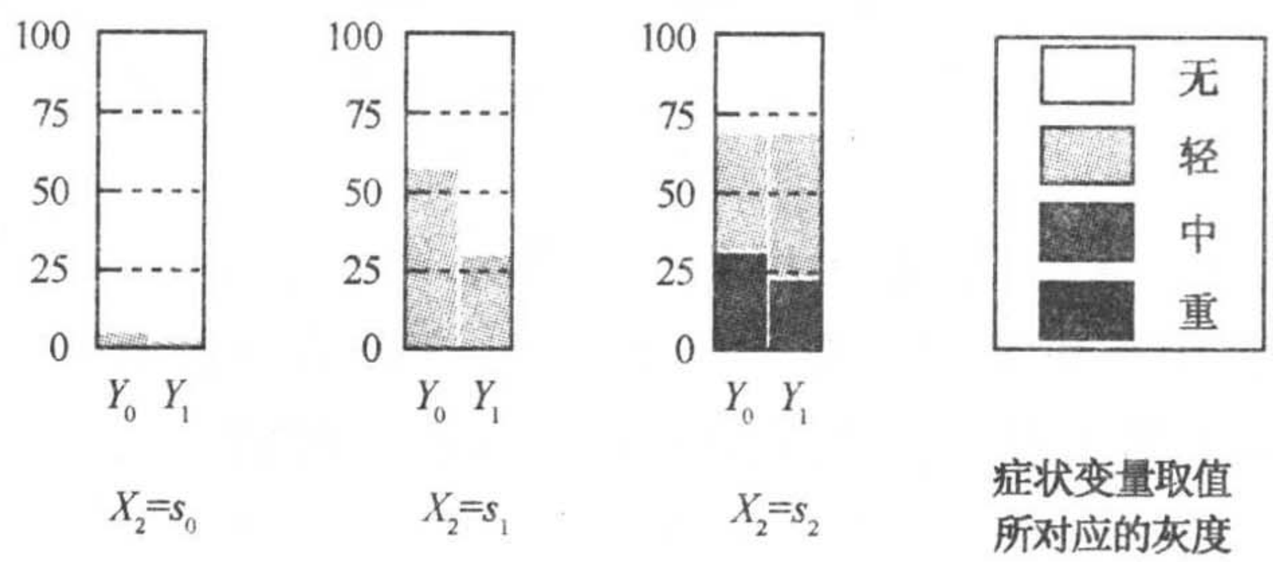


图 10.4　变量 X_2 定义的隐类的类概率直方图，
其中 Y_0 是便稀程度，Y_1 是完谷不化程度

2. 中医理论的定量内容

中医理论没有明确定义证候程度等级，但却隐含了一些有关症状怎样依赖于证候程度变化的内容. 一般地说，证候越重，症状就越多、越重. 对于一个特定症状，证候越重，它出现的可能性就越大，出现的程度就越重. 考虑两个病人甲和乙，如果甲的火不温脾土程度显著重于乙，那么甲出现便稀和完谷不化的概率就会比乙来得大；更确切地说，便稀和完谷不化分别以“重度”、“中或重度”、“轻或中或重度”在甲身上出现的概率都会比在乙身上来得大. 这种关系称为**正单调依赖关系**：便稀和完谷不化正单调依赖于火不温脾土①.

除了症状对证候的正单调依赖关系以外，中医对不同症状出现概率的相对高低，出现时程度的相对轻重也有一定认识. 比方中医认为便稀泻泄多由中焦水谷不分、清浊不别而为之，完谷不化则多由下焦肾虚、命门火衰，水谷不得腐熟蒸化而成，故 前者比后者较易见. 这种关系称为**症状出现相对难易关系**.

中医理论中涉及程度和可能性的内容主要是症状对证候的正单调依赖关系以及不同症状出现的相对难易关系. 下面通过比较会看到，模型 M^* 的内容与这些关系吻合，因此可以说模型 M^* 与中医理论在定量层次相吻合.

3. 比较原理

前面两小节的讨论显示，模型 M^* 的定量内容与中医理论的定量内容很不一样. 它们两者是否可以比较？正单调依赖关系讲述的是当证候轻重程度发生变化时，症状的变化情况；而模型 M^* 则是刻画病人处于某个特定隐类时，或者说证候处于某个特定轻重程度时，症状出现的情况. 前者是动态的，后者是静态的.

---

① 在数学中，一个函数 $f(X)$ 的函数值若随自变量 X 的增大而增大，则称之为正单调函数. 这时，可以说 $f(X)$ 正单调依赖 X.

由于基于对证候处于每一轻重程度的静态刻画，可以推理得到证候在不同轻重程度之间动态变化时的情况，模型 M^* 的定量内容与正单调依赖关系两者之间是可以进行比较的.

作为一个例子，考虑 X_2 所给出的 3 个隐类. 图 10.4 显示，便稀和完谷不化在 $X_2=s_0$ 这类人身上基本不出现；在 $X_2=s_1$ 这类人身上有机会轻度出现；在 $X_2=s_2$ 这类人身上不但有机会轻度出现，而且还有可能中度甚至重度出现. 因此有如下排序：$X_2=s_0$ 最轻，$X_2=s_1$ 较之为重，$X_2=s_2$ 最重. 在这一排序下，Y_0 和 Y_1 明显正单调依赖 X_2：当 X 的取值按照这个序由轻变重时，症状变量 Y_1 和 Y_2 取值为“重”的概率随之增大；它们取值为“重或中”的概率随之增大；它们取值为“重或中或轻”的概率随之增大. 这与中医关于便稀和完谷不化正单调依赖于火不温脾土的论断完全吻合①.

为了反映出上述排序并且方便使用，可以把 X_2 的 3 个取值 s_0，s_1，s_2 分别命名为“无”、“轻”、“重”. 要注意的是，医师心目中有关于火不温脾土各种程度的直观概念，且不同医师的概念可能不同. 这里的“无”、“轻”、“重”不一定是人们心目中的轻重概念的真实写照，它们是对 3 个类的命名，而这 3 个类的统计特性由图 10.4 中的 3 幅直方图所给出.

将症状出现相对难易关系与模型 M^* 进行对比是一项直截了当的工作. 比方，图 10.4 显示，当 $X_2=s_1$ 时，完谷不化出现的概率比便稀小；当 $X_2=s_2$ 时，完谷不化与便稀出现的概率相当，但是便稀中度或重度出现的概率相对为大. 这些恰恰反映了便稀比完谷不化容易见到这一关系.

4. 比较详情

下面逐个考虑 M^* 中隐指标所定义的隐类，将关于它们的定量内容与中医理论的相关内容进行比较，同时对它们的意义进行诠释.

(1) 隐指标 X_1 所定义的隐类

在 10.5 节，隐变量 X_1 被诠释为肾阳虚失温煦程度. 它有 5 个取值 s_0，s_1，s_2，s_3 和 s_4，这即是说模型 M^* 按肾阳虚失温煦程度这个隐指标把人群聚成了 5 个类. 为了把握这 5 类的特性，考虑 X_1 在图 10.2 中所辖症状变量，即便稀 (Y_0)、完谷不化 (Y_1)、畏寒怕冷 (Y_2)、四肢冷 (Y_3) 和腰背冷 (Y_4) 在各个类中的分布情况，如图 10.5 所示. 注意，模型 M^* 的参数包括条件概率分布 $P(Y_2|X_1)$，$P(Y_3|X_1)$ 和 $P(Y_4|X_1)$，但不包括 $P(Y_0|X_1)$ 和 $P(Y_1|X_1)$，它们需要

① 这里先按 Y_0 和 Y_1 的概率分布 把 X_2 定义的 3 个隐类排序，然后指出在该排序下，Y_0 和 Y_1 正单调依赖 X_2. 这似乎是循环推理. 其实不然. 实际上，只要存在一个排序使得 Y_0 和 Y_1 正单调依赖 X_2，都可以得出正文中的结论.

从 $P(Y_0|X_2)$，$P(Y_1|X_2)$ 和 $P(X_2|X_1)$ 中计算出来.

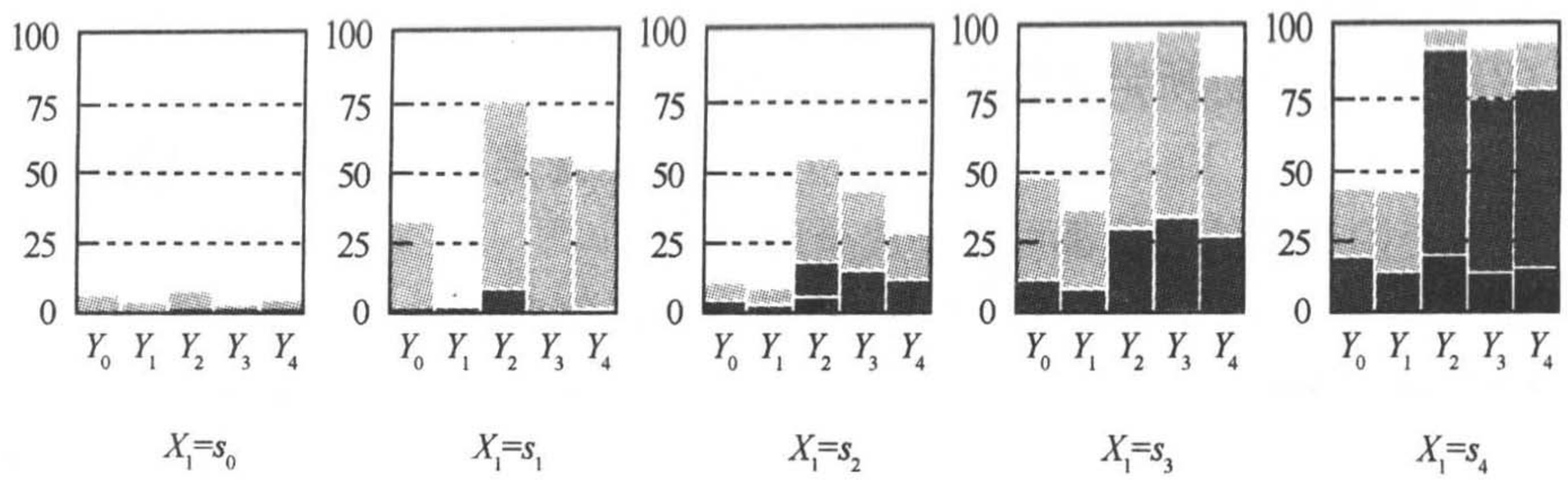


图 10.5　变量 X_1 所定义的隐类的类概率直方图，其中 Y_0 为便稀程度，Y_1 为完谷不化程度，Y_2 为畏寒怕冷程度，Y_3 为四肢冷程度，Y_4 为腰背冷程度

隐变量 X_1 的取值该如何排序呢？图 10.5 显示：$X_1=s_1$ 和 $X_1=s_2$ 比 $X_1=s_0$ 重；$X_1=s_3$ 比 $X_1=s_1$ 和 $X_1=s_2$ 重；$X_1=s_4$ 比 $X_1=s_3$ 重. 但是，$X_1=s_1$ 与 $X_1=s_2$ 之间没有明显的序关系：各症状在类 $X_1=s_1$ 中出现的总概率大于它们在类 $X_1=s_2$ 中出现的总概率；但是，各症状在类 $X_1=s_2$ 中以中度或重度出现的概率却大于它们在类 $X_1=s_1$ 中以中度或重度出现的概率. 所以，可以把 X_1 的取值 s_0，s_1，s_2，s_3 和 s_4 分别命名为“无”、“轻 A”、“轻 B”、“中”和“重”.

注意，在 X_2 取值的排序中，任何两个取值都可以比较轻重，因此它被称为**全序**. 而在 X_1 取值的排序中，有两个取值，即 $X_1=s_1$ 与 $X_1=s_2$，无法比较轻重，因此它被称为**偏序**. 从图 10.5 中不难看出，当 X_1 按这个偏序由轻变重时，即由 s_0 变到 s_1 再变到 s_3，s_4 时，或者由 s_0 变到 s_2 再变到 s_3，s_4 时，症状变量 $Y_0 \sim Y_4$ 取值为“重”的概率、取值为“重或中”的概率以及取值为“重或中或轻”的概率基本上都随之增大①. 所以，便稀等 5 个症状对肾阳虚失温煦的正单调依赖关系与模型 M^* 基本吻合.

在讨论正单调依赖关系时，隐变量取值的排序既可以是全序，也可以是偏序. 读者自然要问，对这个序有没有什么要求呢？回答是，从中医角度看它必须是合理的. 比方，$X_1=s_1$ 与 $X_1=s_2$ 的不可比性就是可以解释的，因为临床实际中的确有两类不同的病人：一类症状相对多而程度轻，另一类症状相对少而程度重，难说谁轻谁重.

就症状出现相对难易关系而言，中医认为阳主外，具有温煦功能，故阳气不足则必以四肢身形失温煦的表现为主，继而阳衰脏寒方可见水谷不腐、清浊不分之症. 因此，畏寒怕冷、四肢冷和腰背冷出现的概率也大于便稀和完谷不化. 图

① 唯一例外是 $P(Y_3$ 为重或中或轻 $|\ X_1=s_3)$ 略大于而不是小于 $P(Y_3$ 为重或中或轻 $|\ X=s_4)$.

10.5 恰恰反映了这一点：无论是在哪一类，便稀和完谷不化出现的概率都远低于其它 3 个症状.

（2）其它隐指标所定义的隐类：全序情况

下面讨论 X_5，X_6，X_7，X_{11} 和 X_{12} 所定义的隐类之统计特征. 根据这些特征，这几个隐变量的取值都可以排成全序.

隐变量 X_5，X_6，X_7，X_{11} 和 X_{12} 的每一个取值对应一个隐类. 图 10.6 展示了各隐变量所辖症状变量在个类中的分布情况. 隐变量 X_5 和 X_{12} 的类概率分布直方图的特点与 X_2 的类概率分布直方图十分相似，因此它们的取值 s_0，s_1 和 s_2 可以分别命名为“无”、“轻”和“重”. 隐变量 X_6 和 X_7 的类概率分布直方图也有与 X_2 的类概率分布直方图相似的特点，不过当它们取值为 s_0 时，相关症状有不小概率会轻度出现，因此把它们的取值 s_0，s_1 和 s_2 分别命名为“轻”、“中”和“重”.

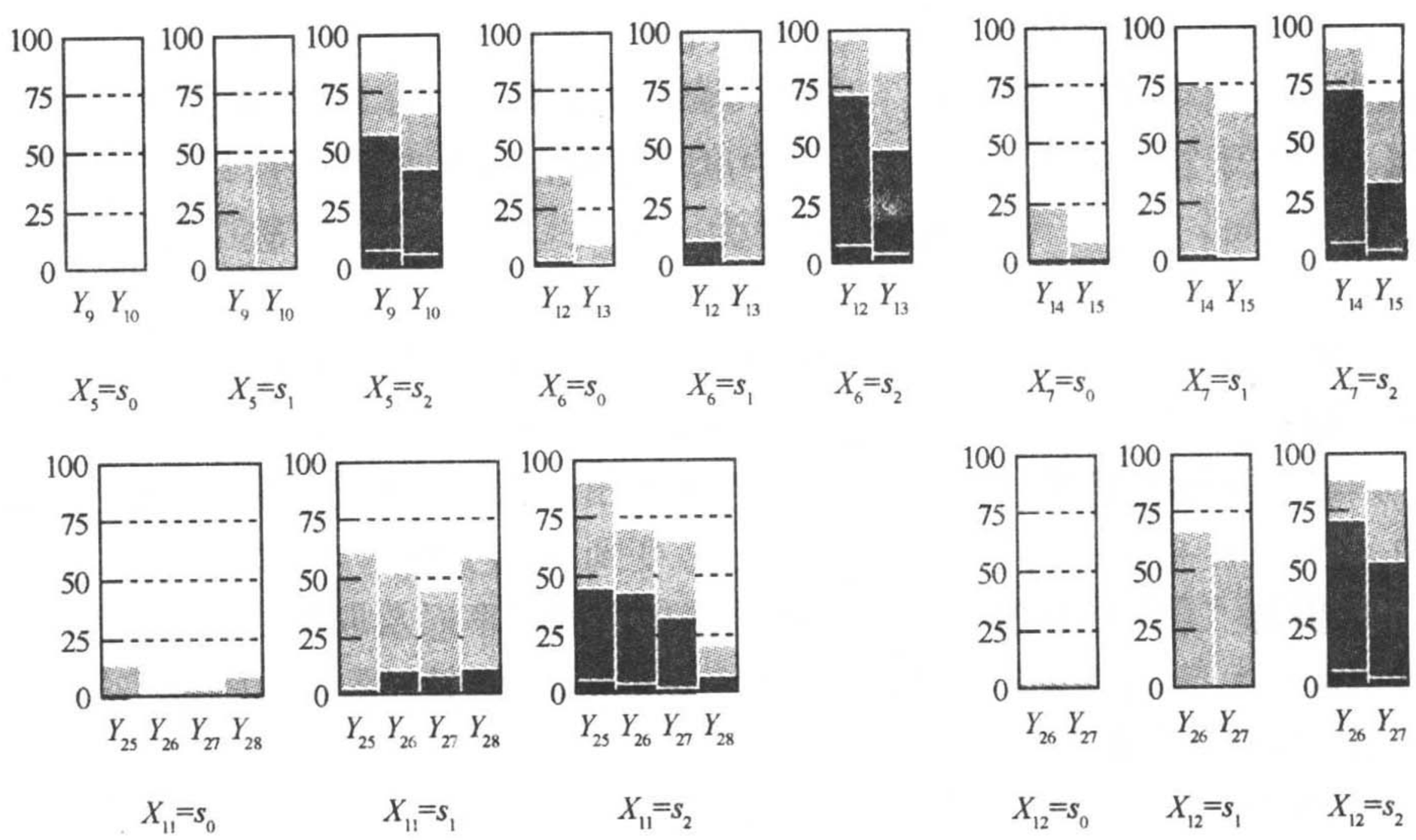


图 10.6　变量 X_5、X_6、X_7、X_{11}、X_{12} 所定义的隐类之类概率直方图

下面考虑隐变量 X_{11}（肾阴虚阳亢程度）所定义的隐类. 在中医中，症状对证候的依赖关系一般是正单调的，这里我们遇到一个例外. 在 X_{11} 所辖的 4 个症状变量中，心慌心跳（Y_{25}）、五心烦热（Y_{26}）、身潮热（Y_{27}）三者正单调依赖肾阴虚阳亢程度. 基于这几个症状得到排序：$X_{11}=s_0$ 最轻、$X_{11}=s_1$ 较之为重、$X_{11}=s_2$ 最重. 例外的是脉细数（Y_{28}）. 我们看到，当 X_{11} 从 s_0 变到 s_1 再变到 s_2 时，脉细数出现的概率先由小变大，然后由大变小. 假若脉细数正单调依赖阴虚

阳亢的话，就会出现矛盾．然而，脉细数对阴虚阳亢的依赖关系正好不是正单调的．实际上，脉细数是脉细加脉数；当阴虚阳亢程度在低水平增强时，脉细和脉数出现的概率都会增加；但当阴虚阳亢增强到一定程度以后，脉象将主要表现为细弱无力，而数反而不明显，于是脉细和脉数同时出现的概率反而降低．模型 M^* 能够反映出这个特别的例外，有力地说明它是高质量的，与中医理论吻合．基于上述讨论，可以把 X_{11} 的取值 s_0，s_1 和 s_2 分别命名为“轻”、“中”和“重”．

（3）其它隐指标所定义的隐类：偏序情况

下面讨论 X_3，X_4，X_8，X_9 和 X_{13} 所定义的隐类之统计特征．根据这些特征，这几个隐变量的取值都只可以排成偏序．

先来看看 X_4，它所辖症状变量在它所定义的各隐类中的分布情况如图 10.7 第一行的 4 幅直方图所示．这些直方图与 X_1 的类概率分布直方图有相似的地方，也有不同之处．相似的地方是类 $X_4=s_1$ 与类 $X_4=s_2$ 不可比较，各症状在类 $X_4=s_1$ 出现的总概率大过它们在类 $X_4=s_2$ 出现的总概率，但是它们在类 $X_4=s_1$ 以重度或中度出现的概率却小于它们在类 $X_4=s_2$ 以重度或中度出现的概率．不同的地方有两点：一是 X_4 只有 4 个取值，而 X_1 则有 5 个取值；二是在类 $X_4=s_0$ 中相关症状有一定概率会轻度出现，而在类 $X_1=s_0$ 中相关症状几乎不出现．基于这些原因，可以把 X_4 的 4 个取值 s_0，s_1，s_2 和 s_3 分别命名为“轻”、“中 A”、“中 B”和“重”．

就症状出现相对难易关系而言，中医认为当肾气亏虚，膀胱气化失司，失去约束固摄功能，轻则夜尿频多、小便余沥不尽，重则昼小便失禁、夜遗尿．这与 X_4 的类概率分布不谋而合：图 10.7 显示，只有在 $X_4=s_3$（“重”）时，昼小便失禁（Y_9）和夜遗尿（Y_{10}）才有较大概率会出现；即使这时，它们重度出现的概率还是很小．另外，中医认为夜尿频多（Y_8）是肾虚的典型症状，这与它在 $X_4=s_1$，$X_4=s_2$ 和 $X_4=s_3$ 这 3 个类中出现的概率都很大的事实一致．

图 10.7 还展示了隐变量 X_3，X_8，X_9 和 X_{13} 所辖症状变量在它们所定义的隐类中的分布情况．我们看到，X_8，X_9 和 X_{13} 的类概率直方图与 X_4 的类概率直方图有相同的特性，因此把它们的取值 s_0，s_1，s_2 和 s_3 分别命名为“轻”、“中 A”、“中 B”和“重”．隐变量 X_3 的类概率直方图也有与 X_4 的类概率直方图相似的特性，但是，在类 $X_3=s_0$ 中，相关症状几乎不会出现．因此，可以把 X_3 的取值 s_0，s_1，s_2 和 s_3 分别命名为“无”、“轻 A”、“轻 B”和“重”．

（4）隐指标 X_{10} 所定义的隐类

下面讨论 X_{10} 所定义的隐之统计特征．基于这些特征，X_{10} 的取值只可排成偏序，而其排成偏序的理由与 X_1 有所不同．

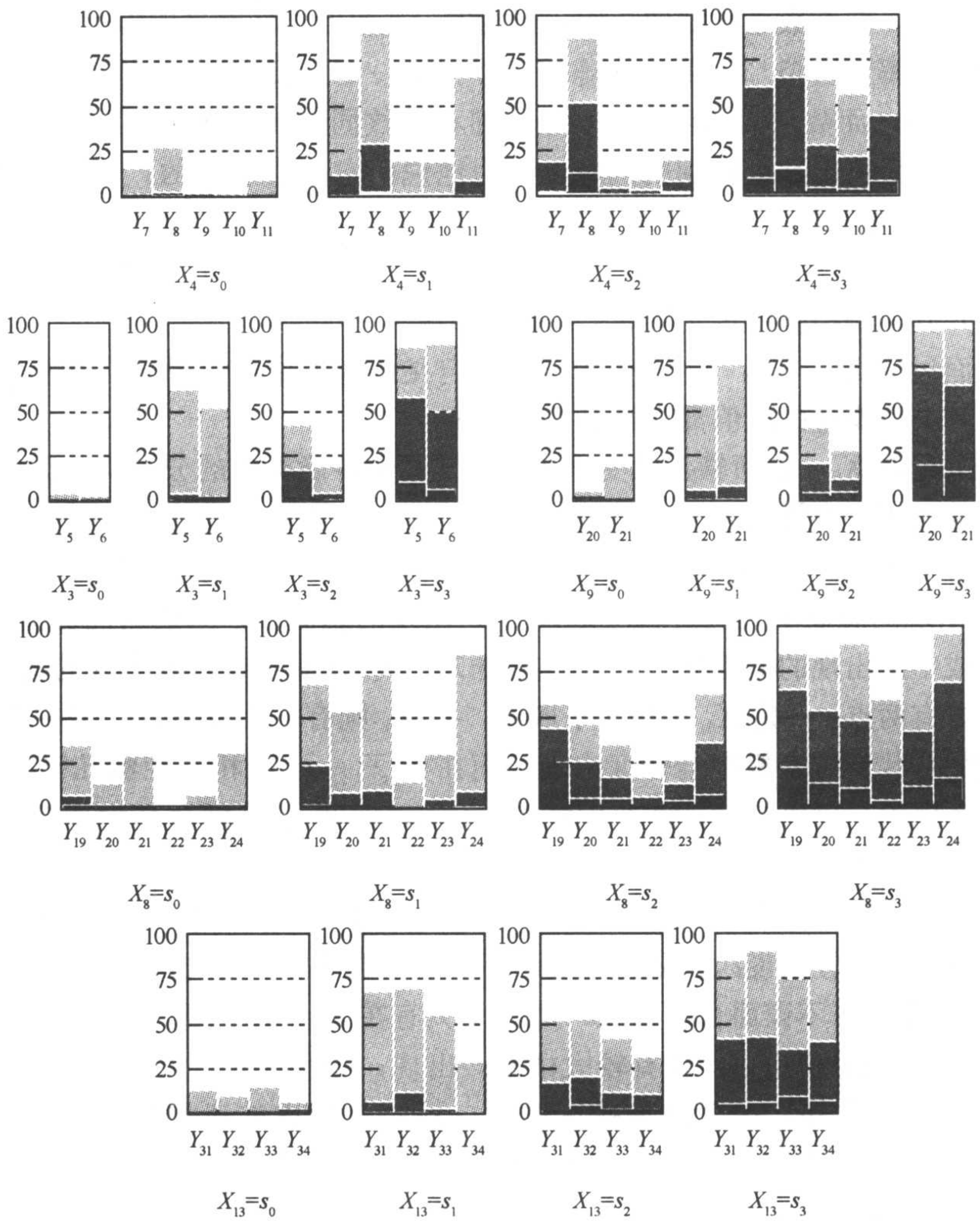


图 10.7　变量 X_3、X_4、X_8、X_9 和 X_{13} 所定义的隐类的类概率直方图

隐变量 X_{10} 所辖症状变量在所定义的各隐类中的分布情况如图 10.8 所示. 由于脉细数（Y_{28}）不正单调依赖于肾阴虚程度，暂时先不考虑它. 其它症状变量的分布情况显示，类 $X_{10}=s_0$ 的病情最轻，类 $X_{10}=s_4$ 的病情最重. 另外的 3 个类之间没有明显的序的关系：在类 $X_{10}=s_1$ 中，失眠（Y_{29}）和头晕（Y_{30}）出现的概率比在其它两类中大；在类 $X_{10}=s_2$ 中，心慌心跳（Y_{25}）、五心烦热（Y_{26}）和身潮热（Y_{27}）出现的概率比在其它两类中大；在类 $X_{10}=s_3$ 中，舌上少津（Y_{31}）、口渴喜饮（Y_{32}）、尿黄赤短少（Y_{33}）和小便涩少（Y_{34}）以重度

或中度出现的概率比在其它两类中大．这些现象从中医角度很好解释．阴虚有 3 个方面的表现，即阴虚失养、阴虚阳亢和阴虚津亏．如果不同的类突出不同的方面，类与类之间就没有序的关系．而我们所遇到的正好是这样的情况：在类 $X_{10}=s_1$ 中，阴虚失养最突出；在类 $X_{10}=s_2$ 中，阴虚阳亢最突出；而类 $X_{10}=s_3$ 中，阴虚津亏最突出．基于上述讨论，可以把 X_{10} 的取值 s_0，s_1，s_2，s_3 和 s_4 分别命名为“轻”、“中 A（偏失养）”、“中 B（偏阳亢）”、“中 C（偏津亏）”和“重”．

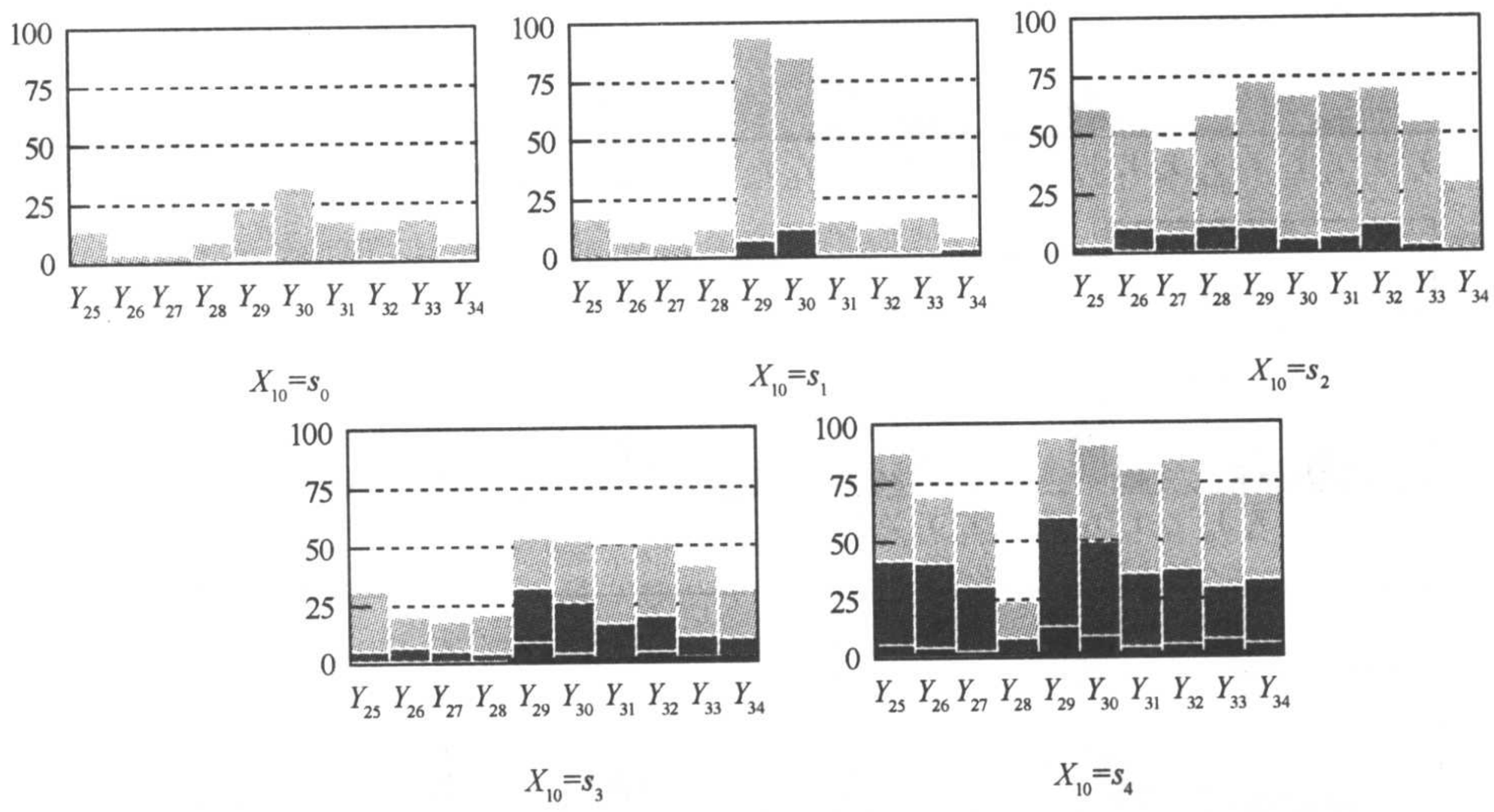


图 10.8　变量 X_{10} 所定义的隐类的类概率分布直方图

回过头来看看脉细数（Y_{28}）的分布．在类 $X_{10}=s_0$ 中，它出现的概率低是因为这类人的病情轻；在类 $X_{10}=s_2$ 中，它出现的概率高是因为这类人的病情中度、偏阳亢；在类 $X_{10}=s_1$ 和 $X_{10}=s_3$ 中，它出现的概率低是因为这两类人的病情中度、偏失养或津亏，阳亢不明显；在类 $X_{10}=s_4$ 中，它出现的概率低是因为这类人阴虚严重，脉象主要表现为细弱无力，而数反而不明显所致．所以，每个情况都是合理的．

（5）隐指标 X_0 所定义的隐类

最后考虑 X_0 这个特殊的隐变量，它的意义是肾虚程度，有 5 个取值，即 s_0，s_1，s_2，s_3 和 s_4．这就是说，模型 M^* 把调查所得到的病例按肾虚程度这个隐指标聚为 5 类．各症状变量在这些类中的分布情况如图 10.9 所示．为了给读者提供另外一个视角，表 10.4 给出了各类中症状出现总数的平均值，或者说是期望值．例如，类 $X_0=s_4$ 中的个体平均有 11.7 个轻度症状、10.2 个中度症状、2.2 个重度症状．

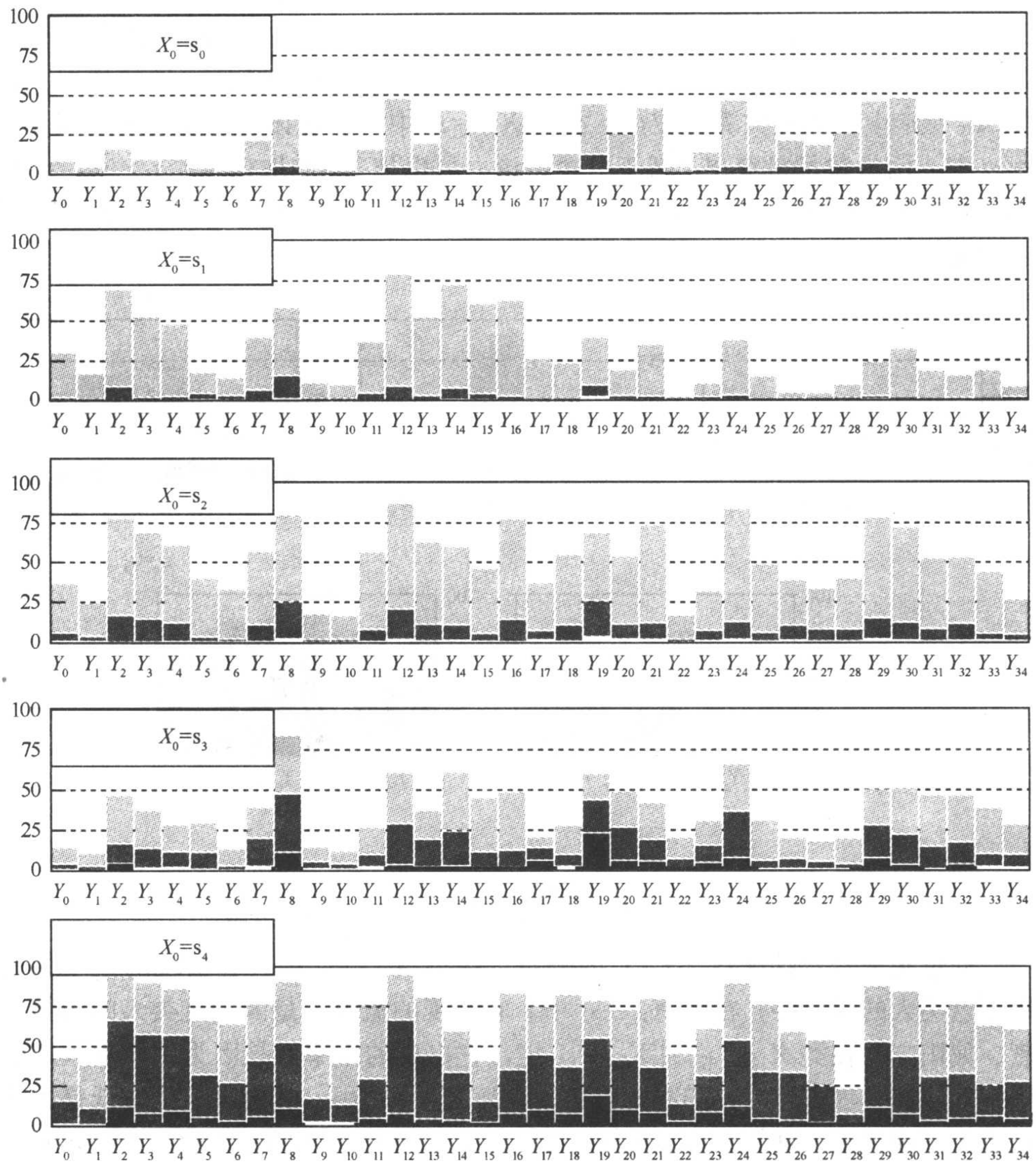


图 10.9　X_0 所定义的隐类的类概率直方图. X_0 的概率分布为 $s_0(0.27)$，$s_1(0.15)$，$s_2(0.21)$，$s_3(0.24)$，$s_4(0.13)$

图 10.9 显示，在类 $X_0=s_0$ 或类 $X_0=s_1$ 中，几乎所有的症状都有或大或小的概率会出现，因此两者均不能解释为“无”. 另一方面，除了诸如畏寒怕冷（Y_2）、夜尿频多（Y_8）、齿松脱落（Y_{19}）和失眠（Y_{29}）等几个症状有少许概率会中度出现以外，其它各种症状出现的程度均是轻度. 因此，s_0 和 s_1 都应该解释为“轻”. 同时它们又是有区别的：在类 $X_0=s_0$ 中，反映阴虚的症状（Y_{25}、

Y_{26}、…、Y_{34})出现的概率相对为大；相反，在类 $X_0=s_1$ 中则是反映阳虚的症状($Y_0,Y_1,\cdots,Y_6$)出现的概率相对为大. 另外，表 10.4 显示，类 $X_0=s_1$ 中的平均症状个数多于类 $X_0=s_0$ 中的平均症状个数. 因此，可以把 s_0 和 s_1 分别命名为“轻 A（偏阴虚)”和“轻 B（偏阳虚)”.

表 10.4　隐变量 X_0 所定义的隐类中症状总数的期望值

症状程度	轻	中	重
$X_0=s_0$	7.2865	0.8729	0.0697
$X_0=s_1$	9.7465	1.0194	0.0779
$X_0=s_2$	14.6014	3.1003	0.2619
$X_0=s_3$	7.3594	4.2443	1.0657
$X_0=s_4$	11.7218	10.1994	2.1756

接下来考虑 X_0 的其它三个取值. 在类 $X_0=s_4$ 中，各种症状出现的概率普遍很大，而且出现时的程度也往往是中或重度；平均症状个数，尤其是中度和重度症状的平均个数，也比其它情况为多. 因此 s_4 应解释为“重”. 其余的两个取值 s_2 和 s_3 所代表的程度明显介于 s_1 和 s_4 之间，所以二者均可解释为“中”. 同时它们又是有区别的：在类 $X_0=s_2$ 中，症状出现的总概率相对较大，但症状是中或重度的概率则相对较小；在类 $X_0=s_3$ 中，症状出现的总概率相对较小，但症状是中或重度的概率则相对较大. 换个角度看，在类 $X_0=s_2$ 中，平均有 14.6 个轻度症状、3.4 个中或重度症状；而在类 $X_0=s_3$ 中，平均则有 7.4 个轻度症状、5.3 个中或重度症状. 所以，$X_0=s_2$ 时症状偏多偏轻，$X_0=s_2$ 时症状偏少偏重. 因此，可以把 s_2 和 s_3 分别命名为“中 A（症状偏多偏轻)”和“中 B（症状偏少偏重)”.

按照肾虚程度这个隐指标，M^* 把病例聚类为“轻 A、”“轻 B”、“中 A”、“中 B”、“重”五类. 下面，从中医理论角度来探讨这一聚类的合理性. 首先注意到，这里不存在肾虚程度为“无”的类别. 也就是说，调查的对象，即 60 岁或 60 岁以上的的老年人，都会或多或少地有肾虚的表现. 这与中医的认识相吻合，其实早在《黄帝内经》中就曾提出，随着“七七”和“八八”的到来，肾中精气必然会逐渐减少并走向衰竭，而出现肾虚证.

第二，考虑各类中久病史（Y_{17}）的情况. 从图 10.9 中可以看到，在“轻 A”类中，只有约 4%的人有轻度久病史（0.5～1 年)；在“轻 B”类中，轻度久病史的人的比例增加至约 27%；在“中 A”类中，不但有约 30%的 人有轻度久病史，而且开始有约 6%的人有中度久病史（2 ～3 年)；在“中 B”类中，有不同程度久病史的总人数降至约 21%，但却开始有约 6%的人有重度久病史（3

年以上)；最后，在“重”的这一类中，有久病史的人的比例为重度（约 10%），中度（约 35%），轻度（约 31%）. 这与中医理论“久病多肾虚”，“它脏之病穷必及肾”的理论相吻合.

第三，无论是在哪一类中，腰膝酸软（Y_{12}）出现的概率相对于其它症状都是最大的，这充分印证了中医学“腰为肾之府”的理论：肾虚腰膝失养则必然见腰膝酸困的症状. 同样，夜尿频多（Y_8）、尺脉不足（Y_{14}）、精神不振（Y_{16}）、齿松脱落（Y_{19}）、耳鸣（Y_{21}）、记忆力减退（Y_{24}）、失眠（Y_{29}）和头部昏晕（Y_{30}）出现的概率大，也分别体现了肾主水司膀胱气化、尺脉以候肾、肾主骨生髓、脑为髓海、肾开窍于耳等中医理论.

第四，便稀完谷不化（Y_0、Y_1）、浮肿（Y_5、Y_6）、尿失禁（Y_9、Y_{10}）、痴呆（Y_{22}）、下肢困乏不便（Y_{23}）和小便涩少（Y_{34}）这几个变量除了在 $X_0=s_4$（“重”）这类中出现较大概率外，在其它几个类中出现的概率均很低. 而中医恰恰认为这些症状要在肾虚程度较重时才会出现：便稀完谷不化是肾虚严重时肾阳不温脾土的征象；浮肿是由于肾虚严重时，肾虚不能温化水湿而导致水湿泛滥而外溢；尿失禁是肾虚严重时，肾的摄纳功能减退，膀胱失约而致尿失禁固的结果；痴呆更是肾精不足，脑窍失养的严重征候；小便涩少也是肾中阴液极度亏虚，尿液化源不足的症状.

综上所述，模型 M^* 在定量层面与肾虚辨证理论基本吻合，其中的隐变量的取值可以按表 10.5 所示诠释.

表 10.5 模型 M^* 中一些隐变量取值的诠释

	X_1	X_3	X_4	X_5	X_6	X_8	X_{10}	X_{11}	X_{13}
s_0	无	无	轻	无	轻	轻	轻	轻	轻
s_1	轻 A	轻 A	中 A	轻	中	中 A	中 A	中	中 A
s_2	轻 B	轻 B	中 B	重	重	中 B	中 B	重	中 B
s_3	中	重	重			重	中 C		重
s_4	重						重		

10.7 结果模型与辨证论治

模型 M^* 是一个贝叶斯网，如果已知其中一些变量的取值，就能计算出其它变量的后验概率分布. 上面两节对 M^* 中的隐变量及其取值进行了诠释. 比方，隐变量 X_{10} 被解释成肾阴虚程度，而它的 5 个取值被分别命名为“轻”、“中 A”、“中 B”、“中 C”、“重”. 对于一个 60 岁或以上的老年人，如果知道他的症状情

况，就可以计算出他有不同程度肾阴虚的概率，从而得出辨证结果.[①] 这种基于 M^* 的辨证称为**模型辨证**.

作为一个例子，考虑表 10.6 所给出的病例 1. 模型辨证不是要计算所有隐变量的后验概率分布，而是要计算出一些有治疗意义的证候隐变量的后验概率分布. 假设隐变量 X_1、X_3、X_4、X_5、X_6、X_8、X_{10}、X_{11} 和 X_{13} 有治疗意义. 那么，病例 1 的模型辨证结果如下，其中程度等级的意义见第 10.6 节。

表 10.6　病例 1

Y_0. 便稀	无	Y_9. 昼小便失禁	无	Y_{18}. 咳嗽喘息	轻	Y_{27}. 身潮热	无
Y_1. 完谷不化	重	Y_{10}. 夜遗尿	无	Y_{19}. 齿松脱落	中	Y_{28}. 脉细数	无
Y_2. 畏寒怕冷	中	Y_{11}. 昼尿频	无	Y_{20}. 听力减退	轻	Y_{29}. 失眠	中
Y_3. 四肢发冷	轻	Y_{12}. 腰膝酸困	中	Y_{21}. 耳鸣	中	Y_{30}. 头部昏晕	中
Y_4. 腰背冷	中	Y_{13}. 腰痛喜按	轻	Y_{22}. 痴呆	轻	Y_{31}. 舌上少津	无
Y_5. 下肢浮肿	轻	Y_{14}. 尺脉不足	无	Y_{23}. 下肢困乏不便	中	Y_{32}. 口渴喜饮	中
Y_6. 面目浮肿	中	Y_{15}. 脉无力	无	Y_{24}. 记忆力减退	中	Y_{33}. 尿黄赤短少	轻
Y_7. 小便余沥不尽	中	Y_{16}. 精神不振	中	Y_{25}. 心慌心烦	轻	Y_{34}. 小便涩少	中
Y_8. 夜尿频	中	Y_{17}. 久患它病史	轻	Y_{26}. 五心烦热	无		

X_1（肾阳虚失温煦）：重（0.619），中（0.374），轻 B（0.007），轻 A（0），无（0）.

X_3（肾阳虚水泛）：重（0.739），轻 B（0.121），轻 A（0.14），无（0）.

X_4（膀胱失约）：重（0.025），中 B（0.779），中 A（0.192），轻（0.004）.

X_5（尿失禁固）：重（0.004），轻（0.072），无（0.924）.

X_6（腰膝失养）：重（0.68），中（0.30），轻（0.02）.

X_8（肾精亏虚）：重（0.984），中 B（0.01），中 A（0.006），轻（0）.

X_{10}（肾阴虚）：重（0.802），中 C（0.195），中 B（0.002），中 A（0.001），轻（0）.

X_{11}（肾阴虚阳亢）：重（0.094），中（0.24），轻（0.666）.

X_{13}（肾阴虚津亏）：重（0.614），中 B（0.386），中 A（0），轻（0）.

用 M^* 来指导辨证的重要意义在于，它既能大大提高辨证的客观性，又能使辨证定量化. 模型辨证之所以客观有两方面的原因：第一是推理过程的客观性，贝叶斯网中的推理完全是按照概率论原则进行，没有半点主观性；第二是模型 M^* 内容的客观性. 模型 M^* 是计算机按照统计学原则对数据进行分析处理的结果，这里的数据只包括症状显变量的取值，而不包括证候隐变量的取值. 当然在

① 由于 M^* 背后的样本空间是 60 岁或以上的老年人，它只能用于这个人群的辨证.

数据收集过程中，对症状轻重程度的判断仍免不了会有主观性，但是由于排除了对证候隐变量的判断，数据中的主观性得以大大减少. 回归分析和判别分析也是从数据出发，建立统计模型以指导辨证，它们的推理过程也是客观的. 但是，由于它们的出发点包括了对证候隐变量主观判断的数据，其客观性不强.

隐变量的取值对应隐类，所以，计算隐变量的后验概率分布就是计算病人属于各隐类的概率. 模型辨证的定量性表现在两个方面：它首先给出病人属于各隐类的概率；同时还如 10.6 节所述的那样给出了每个隐类统计特性.

模型辨证的客观性有助于提升临床辨证的规范化水平，减少人为因素的影响. 模型辨证的定量性则有助于提高辨证论治水平：我们可以事先对证候隐类进行研究，根据其统计特性找到针对它们的适当治疗方案；然后在临床论治时，根据病人属于各隐类的概率，将这些治疗方案进行适当组合，而得到对病人的治疗方案. 这是隐结构法进一步发展需要做的最重要的工作.

抽象地讲，隐结构法通过数据分析，深化我们对证候的认识，从而可以使辨证论治水平得到提高. 正如清代医家叶天士在《临证指南医案》中说："医道在乎识证、立法、用方，此为三大关键，一有草率，不堪司命. 然三者之中，识证尤为紧要."

10.8 模型辨证的质量

模型辨证是客观的、定量的，但并不意味着它是高质量的. 想象这样一个计算机辨证系统，无论病人是谁、病情如何，它的结果永远是"百分之百健康无病". 这个系统是客观的、定量的，但其辨证质量却很差. 所以，有必要考虑模型辨证的质量是否可以信赖的问题.

如果中医辨证有金标准，这个问题容易回答：可以把模型辨证的结果与金标准进行比较，如果吻合，那么模型辨证的质量就没有问题. 遗憾的是中医辨证并无金标准.

长远来讲，模型辨证的质量应该这样来确定：基于病人属于各隐类的概率以及各隐类的统计特征，定出治疗方案，（在保证病人安全的前提下）将之付诸实施，看看效果如何. 如果效果好，说明模型辨证和论治的质量都高. 不过，要进行这样的研究需要很多资源和时间 ，我们暂时还没有条件这样做. 下面通过将它与专家辨证进行比较，初步探讨模型辨证的质量.

模型辨证和专家辨证都是从病人的症状出发，都是要对病人的证型进行推测，不同的只是推理的基础和推理的结果. 所以，可以就这两方面把模型辨证与专家辨证进行比较. 专家辨证推理的基础是肾虚辨证理论，而模型辨证推理的基础是模型 M^* 的内容. 第 10.5 和 10.6 两节已经在定性和定量两个层次把 M^* 的

内容与肾虚辨证理论进行比较，发现两者基本吻合．本节将通过案例分析把模型辨证和专家组辨证进行比较分析，发现两者实质上基本一致．这些表明，模型辨证的质量是可以信赖的．

案例分析的程序如下：首先选择一些病例，请专家对它们进行证候分析并给出辨证结果；然后对收集到的专家意见进行分析，找出其实质内容；接着对模型辨证的特点进行分析；最后把专家辨证与模型辨证进行比较．

1．案例选择

案例分析选用了 4 份案例，其中的一份见表 10.6，另外 3 份见图 10.11．它们均未参与模型学习．之所以选择这 4 个病例是因为它们充分反映了肾的阴阳虚实变化，从而具有代表性．事实上，病例 1（表 10.6）的特点是症状多，阳虚、阴虚症状都有；病例 2（图 10.11）中基本没有阳虚症状；病例 3（图 10.11）中基本没有阴虚症状；病例 4（图 10.11）中的症状为数不多，阳虚、阴虚症状基本上都没有．

仅仅 4 个案例是否为数太少？案例分析的目的是要揭示模型辨证与专家辨证在哪些方面一致，在哪些方面不同，以及原因何在．4 个案例虽然不多，但是因为它们具有很好的代表性，所以已经基本上可以满足案例分析的需要．另外，篇幅限制也不允许讨论太多案例．

2．专家辨证收集

有 7 位从事中医诊断教学和研究多年并且具有高级职称的专家参与了案例分析．专家们被要求 对上述 4 个病例进行证候分析和辨证．在此过程中，不在表中的变量的取值视为“未知”而不是“无”，而“无”、“轻”、“中”、“重”的意义以 10.2 节的标准为准．由于有 7 位专家和 4 个病例，所以总共有 28 份证候分析．下面是三位专家对病例 1 所做出的证候分析和辨证结论：

(1) 专家甲：

辨证结果：肾气不固证(中度)，肾阳虚证(中度)，兼有肾精不足(轻度)．

证候分析：年高肾虚，机能活动减退，则腰膝酸困、耳鸣、精神不振；肾与膀胱相表里，肾气虚，膀胱失约，以致夜尿频多，排尿机能无力，尿液不能全部排出，可使尿后余沥不尽；肾阳亏虚，不能温煦肌肤，则畏寒怕冷、腰背冷、四肢冷；命门火衰，火不生土，脾失健运，故久泻不止、完谷不化；肾津不足，则口渴喜饮．

(2) 专家乙：

辨证结果：肾虚证（重度）．

证候分析：腰为肾府，肾主骨，肾虚不能温养骨骼，可见腰膝酸困、腰痛喜按；命门火衰，火不生土，脾失健运，完谷不化；阳虚

失煦，则畏寒怕冷、四肢冷、腰背冷；阳气不足，心神无力振奋，则精神不振；阳虚水泛，则下肢浮肿、面目浮肿；上逆犯肺，可见咳嗽喘息；肾精不足，导致齿松脱落、听力减退、耳鸣、痴呆、记忆力减退；肾阴虚，可见心慌心烦、失眠、头部昏晕、口渴喜饮、尿黄赤短少、小便涩少.

(3) 专家丙：

辨证结果：肾阳虚证（中度），兼有肾阴虚证（轻度）.

证候分析：肾阳亏虚，脾失温煦，水谷不化，则见完谷不化；肾阳亏虚，虚寒内生，可见畏寒怕冷、腰背冷；肾阳亏虚，水液不化，可见下肢浮肿、面目浮肿、夜尿频多；肾虚是以肾精亏虚为基础，肾精亏虚，不能主骨生髓，可见腰膝酸困、耳鸣、记忆力减退、头部昏晕、精神不振、齿松脱落、痴呆、听力减退；肾阴亏虚，心失滋养，可见心慌心烦、失眠；肾阴亏虚，化源不足，可见口渴喜饮、尿黄赤短少. 肾亏累及它脏，病势尚轻.

3. 专家辨证的特点

专家们的证候分析和辨证结论有相同的地方，也有不同之处. 下面对这些异同进行分析，从而揭示专家辨证的一些特点.

专家证候分析的不同表现在 3 个方面. 首先，专家在辨证时往往会忽略一些症状，而不同专家往往忽略不同症状. 在对病例 1 的分析中，专家甲没有提及下肢浮肿、面部浮肿、心慌心烦、失眠等症状；专家乙没有提及夜尿频多、小便后余沥不尽等症状；而专家丙则没有提及腰痛喜按、四肢冷等症状. 其次，不同专家对个别症状的解释有所不同. 比方，专家乙认为精神不振是阳气不足的表现，而专家丙则认为它是肾精亏虚的结果. 又如，专家甲认为夜尿频多是膀胱失约所致，而专家丙则认为它是肾阳虚水液不化的结果. 第三，不同专家在谈论隐因子时措辞往往不同. 例如在谈论畏寒怕冷的原因时，专家甲说："肾阳亏虚，不能温煦肌肤，则畏寒怕冷、腰背冷、四肢冷"；专家乙说："阳虚失煦，则畏寒怕冷、四肢冷、腰背冷"；而专家丙则说："肾阳亏虚，虚寒内生，可见畏寒怕冷、腰背冷". 虽然表面上看来有所不同，3 位专家其实都是在说：有一个隐因子，它导致畏寒怕冷、腰背冷、四肢冷. 不同的只是对这个隐因子的称谓.

措辞的不同是证候分析因人而异的主要原因. 如果排除措辞的影响，用统一的词汇，3 位专家的证候分析所揭示的隐因子分别为

(1) 专家甲：肾虚腰膝失养（机能活动减退，则腰膝酸困）[①]，肾阳虚火不

① 括号前是统一词汇，括号里是专家原文.

温脾土（火不生土，脾失健运），肾阳虚失温煦（肾阳亏虚，不能温煦肌肤），肾气虚膀胱失约，肾精不足（机能活动减退，则耳鸣、精神不振），肾阴虚津亏（肾津不足，则口渴喜饮）.

（2）专家乙：肾虚腰膝失养（肾虚不能温养骨骼，可见腰膝酸困、腰痛喜按），肾阳虚火不温脾土（火不生土，脾失健运），肾阳虚失温煦（阳虚失煦），肾阳虚水泛（阳虚水泛，上逆犯肺，可见咳嗽喘息），肾气虚（阳气不足），肾精不足，肾阴虚失滋养（肾阴虚，可见心慌心烦、失眠、头部昏晕），肾阴虚津亏（肾阴虚，可见口渴喜饮、尿黄赤短少、小便涩少）.

（3）专家丙：肾虚腰膝失养（肾精亏虚，不能主骨生髓，可见腰膝酸困），肾阳虚火不温脾土（肾阳亏虚，脾失温煦），肾阳虚失温煦（肾阳亏虚，虚寒内生），肾阳虚水泛（肾阳亏虚，水液不化），肾精不足（肾精亏虚），肾阴虚失滋养（肾阴亏虚，心失滋养），肾阴虚津亏（肾阴亏虚，化源不足）.

上面的分析显示，尽管 3 位专家的证候分析表面上有许多不同，但实质上并没有太大差别.

3 位专家的证候分析实质上是基本一致的，但他们的辨证结论却有较为显著的差别. 这是为什么？原因可能是，在从证候分析到辨证结论的过程中已经加入了论治的因素，专家在决定辨证结论时已经考虑到了治疗原则. 由于临床背景经验的差别，不同的专家可能习惯于不同的治疗取向，因此基于相同或类似的证候分析，有可能得出不同的辨证结论. 专家丙的结论是“中度肾阳虚证兼轻度肾阴虚证”，表明他在治疗上会着重温补肾阳，兼适当地滋肾阴. 专家甲强调肾气不固和肾阳虚，他虽然会同专家丙一样用温补肾阳之药，但他认为首先要解决的是肾气不固，用药可能会以固摄药物为主，辅以适当温阳和滋阴之品. 而专家乙的结论则是“重度肾虚”，表明他会选用一般的补肾药物，而不会特别突出温阳或滋阴.

4. 模型辨证的特点

在上一节我们看到，模型辨证是相对于一些证候隐变量对病人的病理状态做出判断. 它与专家辨证有一个重要的不同之处，那就是它不考虑治疗. 因此，在功用上它相当于证候分析，而不是辨证结论.

5. 专家证候分析与模型辨证的比较

专家证候分析与模型辨证在内容上有很大区别. 专家证候分析指出证候因子是否在病人身上出现，而模型辨证则给出证候因子在病人身上出现的概率分布. 因此，将它们两者进行比较有一定困难. 但是，比较仍然是可能的. 模型辨证是对关于证候因子的不确定性进行定量刻画；而专家证候分析则是按不确定性大小

对证候因子进行选择. 因此，如果“若模型辨证指出证候因子 A 出现的概率及程度大于证候因子 B 出现的概率及程度，则 A 被专家选择的可能性就比 B 大；如果有众多专家，那么选择 A 的人就会比选择 B 的人多.”的关系成立，则可以认为两者吻合.

下面把这个关系作为比较模型辨证和专家证候分析的基础.

病例 1 的模型辨证结果已在 10.7 节给出. 各证候因子被专家选择的次数如表 10.7 所示. 表中的数字是通过对每位专家的意见进行如 10.8.3 节的分析，然后进行统计而获得的. 有一点需要特别注意：肾阴虚分为肾阴虚阳亢、肾阴虚津亏、肾阴虚失滋养 3 个方面. M^* 中的隐变量 X_{11}，X_{13}对应前面两个方面，但是没有隐变量对应第三个方面. 第三方面的症状，即失眠和头部昏晕，与隐变量 X_{10}相连. 所以选用 X_{10}作为比较对象之一，尽管它与 X_{11}和 X_{13}在覆盖范围上有些重复.

表 10.7 病例 1 的各证候因子被专家选择的次数

证候因子	选择专家比例	证候因子	选择专家比例
X_1（肾阳虚失温煦）	6/7	X_{10}（肾阴虚）	7/7
X_3（肾阳虚水泛）	6/7	X_{11}（肾阴虚阳亢）	1/7
X_4（膀胱失约）	3/7	X_{13}（肾阴虚津亏）	5/7
X_5（尿失禁固）	0/7	肾阴虚失滋养	4/7
X_6（腰膝失养）	5/7	水气凌心	2/7
X_8（肾精亏虚）	7/7		

下面就病例 1 把模型辨证与专家们的集体意进行比较. 为了一目了然，我们把模型辨证所给出的概率分布和各证候因子被专家选择的情况用直方图表示出来，见图 10.10 第一排左、中两图. 在模型辨证直方图中，柱体不同灰度段的长度代表证候隐变量取不同值的概率大小，其中灰度与变量取值的对应由第二行的图例给出. 在专家辨证直方图中，柱体的高度代表选择各证候因子的专家在专家群体中占的比例. 模型辨证与专家们的集体意见十分之吻合，这可以从 5 个方面来看：

(1) 根据模型辨证，病人有很大可能性（98%、80%）有重度肾精亏虚（X_8）和肾阴虚（X_{10}），而 7 位专家全部选择了这两个证候因子；

(2) 根据模型辨证，病人有较大可能性（62%、74%、68%、61%）有重度肾阳虚失温煦（X_1）、肾阳虚水泛（X_3）、腰膝失养（X_6）、肾阴虚津亏（X_{13}），而 7 位专家中有 5 至 6 位选择了这些证候因子；

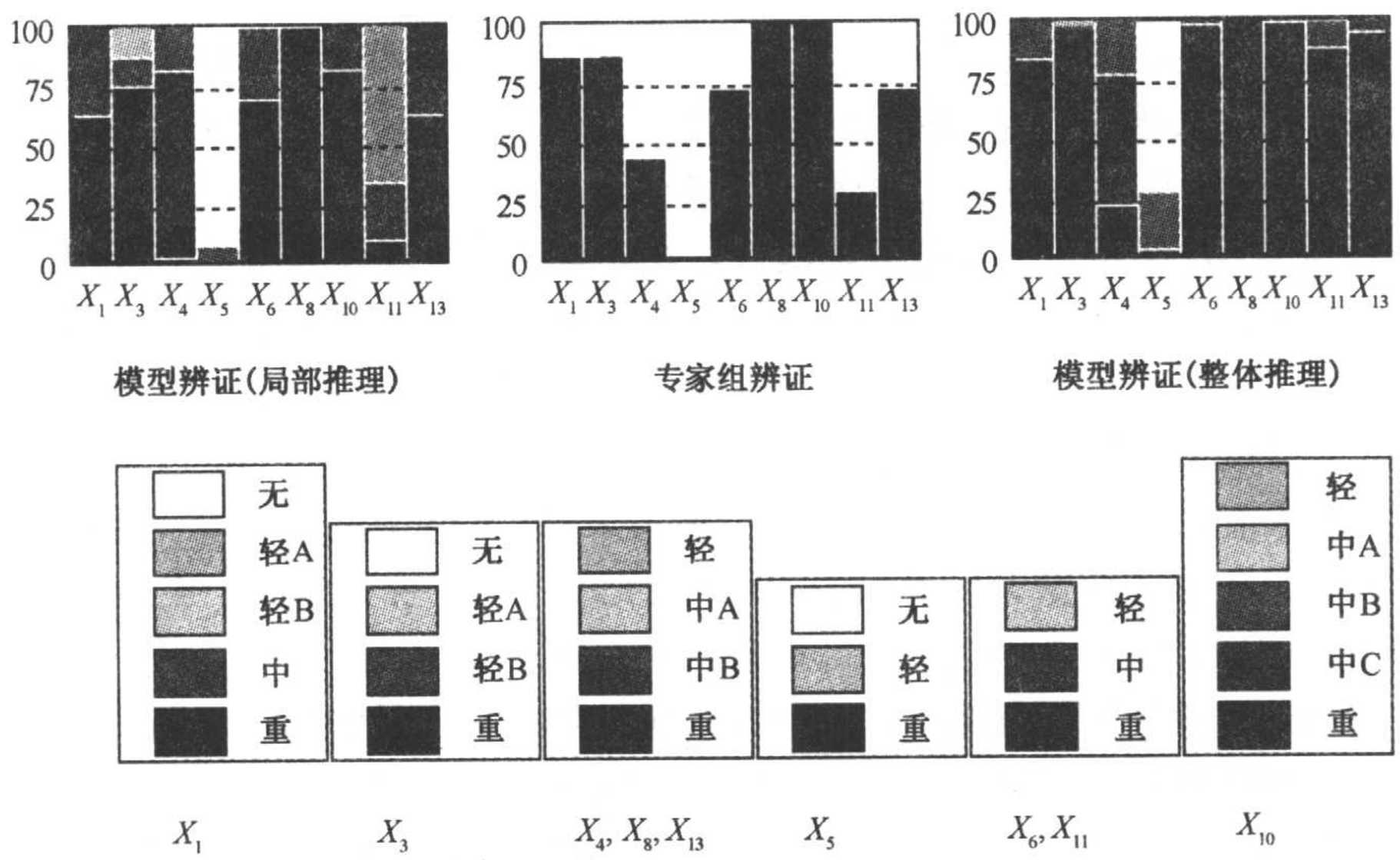


图 10.10　模型辨证与专家证候分析在病例 1 上的比较

(3) 根据模型辨证，病人有很大可能性（97%）有中度膀胱失约（x_4），而 7 位专家中 3 位选择了这个证候因子；

(4) 根据模型辨证，病人有中度或轻度肾阴虚阳亢（X_{11}）的概率分别是 24%和 67%，而 7 位专家中有 2 位选择了这个证候因子；

(5) 根据模型辨证，病人有尿失禁固（X_5）的可能性很小（1%），而 7 位专家中没有 1 位选择这个证候因子.

对比的结果也有不十分吻合的地方. 7 位专家中有 3 位明确指出了肾阴虚失滋养这个证候因子①，有 2 位提到水气凌心. 由于 M^* 中没有与之对应的隐变量，这两点在模型辨证中没有得到很好反映.

6. 局部推理与全局推理

整体观是中医的基本特点. 在中医理论中，不同证候之间存在着广泛联系. 比方，肾阴虚可以引起肾阳虚，肾阳虚也可能导致肾阴虚，等等. 然而，在证候分析时，医师却往往只考虑局部因素：在判断病人是否有某证候时，医师往往只看该证候的症状是否出现，而不会特别在意其它症状. 从前面给出的证候分析例子中可以看出，在判断病人是否有肾阳虚失温煦时，专家们考虑的只是病人是否

① 有一位专家选择的是肾阴虚. 由于肾阴虚失滋养是肾阴虚的一个方面，所以在表 10.7 中，选择肾阴虚失滋养的有 4 人.

畏寒怕冷、腰背冷、四肢冷；在判定肾阳虚水泛时，他们只提及面目浮肿、下肢浮肿. 他们所用的其实是如下推理方式：因为病人畏寒怕冷、腰背冷、四肢冷，所以是肾阳虚失温煦；因为病人面目浮肿、下肢浮肿，所以是肾阳虚水泛. 推理过程不涉及与所关心证候不直接相关的症状，故可以称其为**局部推理**.

在模型 M^* 中也可以进行**局部推理**：在计算一个证候隐变量的后验概率分布时，只考虑在 M^* 的结构（见图 10.2）中它所辖的症状变量的取值，而不管其它症状变量的取值，即把它们视为未知. 例如，在计算隐变量 X_3（肾阳虚水泛程度）的后验概率分布时，只考虑症状变量 Y_5（下肢浮肿程度）和 Y_6（面目浮肿程度）的取值，而把 $Y_0 \sim Y_4$ 和 $Y_7 \sim Y_{34}$ 视为未知；在计算隐变量 X_1 的后验概率分布时，只考虑显变量 $Y_0 \sim Y_4$ 的取值，而把 $Y_5 \sim Y_{34}$ 视为未知. 上两小节所讨论的模型辨证结果就是用这种局部推理而获得的.

与局部推理不同，**整体推理**同时考虑所有症状显变量的取值. 图 10.10 的右图给出了整体推理就病例 1 所得到的结果. 我们看到，这些结果与局部推理的结果有显著不同，各证候隐变量取值为“重”的概率明显相对变大. 什么原因呢？这是因为证候隐变量之间相互关联之故. 举例来说，在只考虑下肢浮肿、面目浮肿两症状的情况下，已经得到肾阳虚水泛为“重”的概率是 74%的结论；然后进一步考虑诸如畏寒怕冷、腰背冷等肾阳虚失温煦症状，肾阳虚水泛为“重”的概率自然会升高.

既然局部推理与整体推理的结果有显著不同，自然就要问在它们之间如何选择？这个问题其实并不重要，重要的是充分理解它们不同的特点，并且在论治时给以充分考虑. 上面的讨论显示，局部推理与医师的思维习惯是一致的，所以使用起来会方便一些. 另一方面，整体推理考虑问题更全面，它的结果综合性更强，因此它将来或许会对改进证论治辨发挥积极作用.

7. 其它病例

10.8.6 节的讨论显示，只有在采用局部推理时，模型辨证才可以与专家证候分析对比. 所以，以后在提及模型辨证时，都是指采用局部推理的模型辨证. 前面已经就病例 1 把模型辨证和专家证候分析进行了比较. 这一小节来看一看它们两者在其它 3 个病例上的对比情况. 这 3 个病例所包含的症状见图 10.11. 图 10.11还给出了模型辨证结果以及专家意见统计.

就病例 2 而言，模型辨证与专家组的集体意见是十分吻合的. 这一点可以从三个侧面来看. 首先，根据模型辨证，病人有重或中度腰膝失养（X_6）、肾精亏虚（X_8）、肾阴虚（X_{10}）、肾阴虚阳亢（X_{11}）、肾阴虚津亏（X_{13}）的概率几乎都是 100%；在专家组方面，几乎所有专家都选择了所有这些证候因子. 其次，根据模型辨证，病人有中度膀胱失约（X_4）的概率大约是 40%；在专家组方面，7

位专家中有一位选择了这个证候因子. 第三，根据模型辨证，病人不可能有重或中度肾阳虚失温煦（X_1）、肾阳虚水泛（X_3）、尿失禁固（X_5）；在专家组方面，没有专家选这些证候因子.

病例	2	3	4	病例	2	3	4	病例	2	3	4	病例	2	3	4
Y_0	无	无	无	Y_9	无	重	无	Y_{18}	无	无	无	Y_{27}	轻	无	无
Y_1	无	无	无	Y_{10}	无	中	无	Y_{19}	重	轻	重	Y_{28}	中	无	无
Y_2	无	中	无	Y_{11}	无	无	无	Y_{20}	轻	无	轻	Y_{29}	轻	无	重
Y_3	无	无	无	Y_{12}	中	中	无	Y_{21}	无	无	无	Y_{30}	中	轻	中
Y_4	无	中	无	Y_{13}	中	中	无	Y_{22}	无	中	轻	Y_{31}	中	轻	无
Y_5	无	轻	无	Y_{14}	轻	轻	轻	Y_{23}	中	中	中	Y_{32}	轻	无	无
Y_6	无	轻	无	Y_{15}	无	无	无	Y_{24}	中	无	中	Y_{33}	无	无	无
Y_7	无	无	中	Y_{16}	中	重	中	Y_{25}	中	无	无	Y_{34}	轻	无	无
Y_8	轻	无	重	Y_{17}	轻	中	轻	Y_{26}	中	无	无				

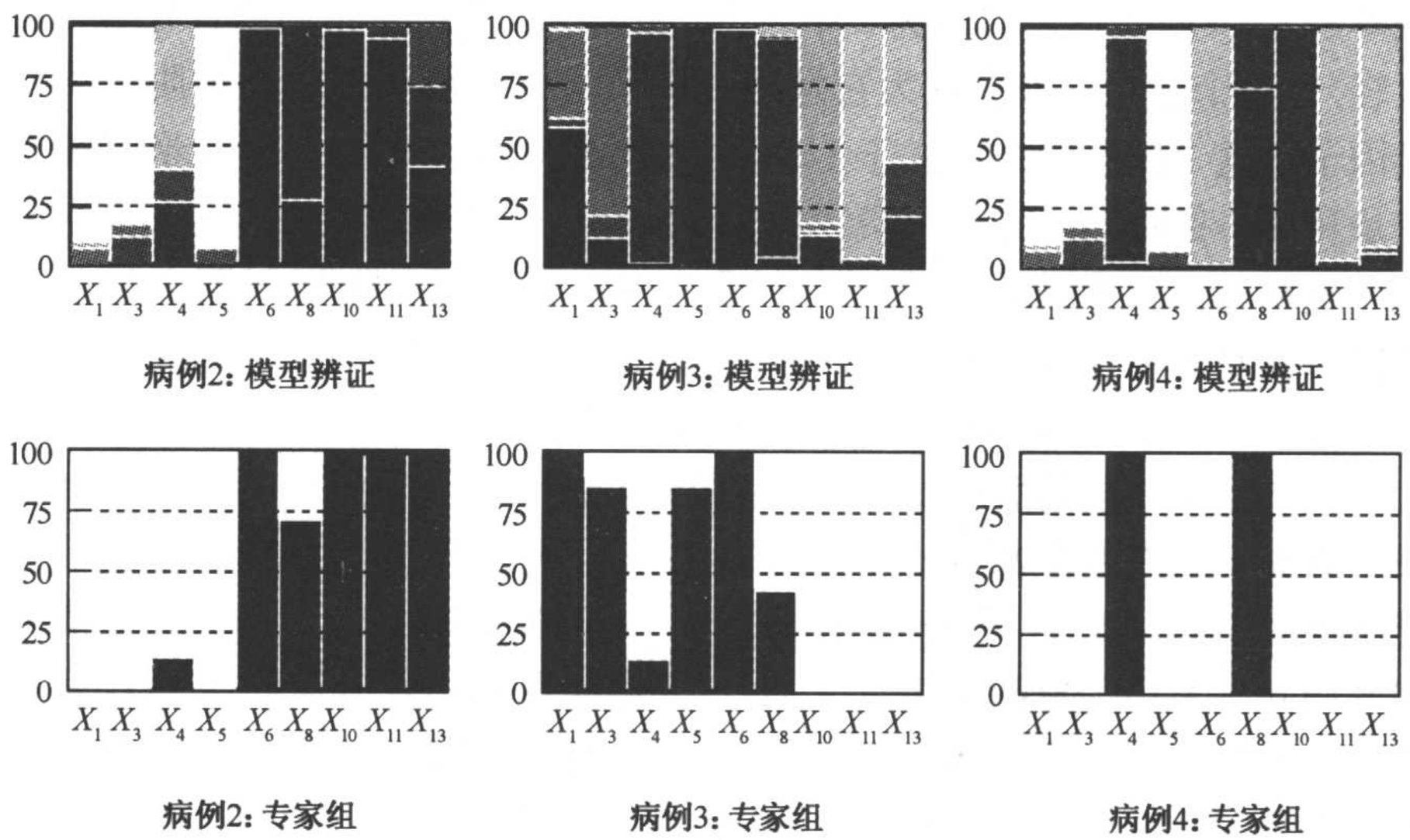


图 10.11　模型辨证与专家证候分析在 3 个病例上的比较

就病例 3 而言，模型辨证与专家组的集体意见大体一致，但也有不吻合的地方. 先来看看两者一致的方面. 首先，根据模型辨证，病人有重度或中度尿失禁固（X_5）、腰膝失养（X_6）的概率分别是 100%和 99.5%；在专家组方面，几乎所有专家都选择了这些证候因子. 其次，根据模型辨证，病人有中度膀胱失约（X_4）、肾精亏虚（X_8）的概率几乎是 100%；在专家组方面，7 位专家中分别有 1 位和 3 位选择了这两个证候因子. 第三，根据模型辨证，病人有重或中度、肾阴虚（X_{10}）、肾阴虚阳亢（X_{11}）的概率甚微；在专家组方面，没有专家选这些

证候因子.

模型辨证与专家组的集体意见在病例 3 上的不吻合表现在肾阳虚水泛（X_3）和肾阴虚津亏（X_{13}）两个证型上. 模型辨证认为，病人有重度肾阳虚水泛（X_3）① 的概率不足 15%，然而 7 位专家中有 6 位选择了这个证候因子. 可能原因有两个：

(1)“下肢浮肿（轻）”和“面目浮肿（轻）”这两个水泛症状位于病例前端，因而受到特别关注；

(2) 专家们没有充分考虑到这两个症状的程度. 类似的问题在肾阳虚失温煦（X_1）上也有发生. 根据模型辨证，病人有重或中度肾阳虚失温煦（X_1）的概率仅仅是 61%，但是所有专家都选择这个证候因子. 这可能与失温煦症状“畏寒怕冷（中）”和“腰背冷（中）”位于病例症状排列前端有关.

模型辨证与专家组的集体意见在肾阴虚津亏（X_{13}）上的不吻合表现如下：由于有“舌上少津（轻）”这个症状，模型辨证得出病人有重或中度肾阴虚津亏的概率是 42%的结论，与病例 2 中之 X_4 相当. 然而，专家们或者忽略“舌上少津（轻）”这个症状，或者认为它是阳虚所致，因此没有人选择肾阴虚津亏这个证候因子. 这个问题的根源在于 HLC 模型的局限性，即一个显变量不能与多个隐变量相连. 变量 Y_{31}（舌上少津程度）已经与 X_{13} 相连，就不能再与其它证候因子隐变量相连. 这意味着，在 M^* 中舌上少津的出现只能用肾阴虚津亏（X_{13}）来解释，其它可能性完全被排除. 当专家引用其它解释时，就出现不吻合.

就病例 4 而言，模型辨证与专家组的集体意见也是大体一致. 根据模型辨证，病人有重或中度膀胱失约（X_4）、肾精亏虚（X_8）的概率几乎是 100%；在专家组方面，所有专家都选择了这两个证候因子. 模型辨证认为，病人有重或中度肾阳虚失温煦（X_1）、肾阳虚水泛（X_3）、尿失禁锢（X_5）、腰膝失养（X_6）、肾阴虚阳亢（X_{11}）、肾阴虚津亏（X_{13}）的概率甚微；在专家组方面，完全没有人选这些证候因子. 关于 X_{10}，模型辨证与专家组的集体意见不是十分吻合. 由于有“失眠（重）”、“头部昏晕（中）”这两个症状，模型辨证认为病人有中度肾阴虚的概率是 100%. 然而，专家们把“失眠（重）”和“头部昏晕（中）”视为肾精亏虚的结果，因而没有一人选择肾阴虚. 与病例 3 中的 X_{13} 的情况相似，这里不吻合的原因在于 HLC 模型的局限性.

10.9 讨　　论

本章介绍了一种研究中医辨证的新方法，即隐结构法. 它的目的是要利用统

① 注意，X_3 的 4 个取值是“重”、“轻 B”、“轻 A”、“无”，它没有“中”这个取值.

计学方法为中医辨证建立客观的、定量的标准. 其基本步骤是：首先对症状在人群中的分布进行流行病学调查，然后利用电脑按照统计学原则对数据进行分析，建立隐结构模型，最后用隐结构模型指导辨证，建立起辨证标准.

隐结构法的核心思想可以归纳为两个基本点和一个中心. 第一个基本点是：中医辨证把病人分成不同的类，自然要问，这些类是否客观存在？如果要从统计学角度回答这个问题，就需要用聚类分析. 第二个基本点是：中医辨证所进行的是多维同时分类，因此简单聚类分析是不够的，需要多维同时聚类；多维聚类中的每个维就是一个隐指标，所以需要隐结构模型. 一个中心是：我们的祖先已经构造了一个隐结构模型，那就是中医理论；由于它是由人脑构造，这个模型缺乏客观性、定量性；隐结构法的基本思想是利用电脑来构造隐结构模型.

隐结构法与以前的方法有本质区别：它的出发点是通过望、闻、问、切这些传统手段所获得的病情资料，不像证实质研究那样需要引用实验室指标，也不像回归/判别分析那样需要依赖专家做出关于证型的主观判断. 它与中医理论的自然形成及其应用类似：两者都是始于对症状在不同个体身上出现情况的观察，通过观察发现规律，为了解释规律引进隐结构模型，最后用隐结构模型指导辨证. 所不同的是中医所依赖的观察是不系统的，处理观察资料用的是人脑，所得到的隐结构模型是用自然语言描述的，缺乏定量性、客观性；而隐结构法所依赖的数据是通过系统流行病学调查的结果，处理数据用的是电脑加统计学原则，所得到的隐结构模型是用数学语言描述的，具有定量性、客观性.

隐结构法实际上是一个基本框架，在这个框架下，随着所使用的模型以及相关算法的不同，有许许多多不同的具体方法. 隐结构法是否可行要看它是否可以帮助我们获得与中医实际情况相符合的模型. 为了探索隐结构法的可行性，张连文、袁世宏（2004）用 HLC 模型对老年人群中的肾虚辨证进行了研究. 所得到的模型无论是在定性层面还是在定量层面都与肾虚辨证理论基本吻合，而且用它来指导辨证在一组具有代表性的案例上所得的结论与几位专家的集体意见实质上一致. 这些都显示隐结构法是可行的.

将来的研究有两个主要方向. 第一个方向是把隐结构法应用于中医辨证各个不同领域以及不同人群的研究. 张和袁用隐结构法对肾虚辨证进行了初步研究，自然也可以用它来研究辨证的其它子领域，比方肝病辨证、胃病辨证，以及八纲辨证、气血津液辨证等. 隐结构法还可以用于中西医结合研究. 辨证施治是中医的特点，辨病治疗是西医的原则. 中西医结合首先辨病分期，然后进行中医辨证分型、辨证施治. 这里的一个关键问题是：一个西医疾病（比方高血压）中有几个中医证型，它们之间如何区别？隐结构法可以用来为揭示西医疾病的中医证候结构，深化对它的认识，建立客观定量辨证标准，从而提高辨证论治水平.

将来的另一个主要研究方向是关于隐结构分析工具的研究. 目前 Zhang 和

Kocka（2004b）的启发式单重爬山法是学习 HLC 模型最好的算法. 但是，这个算法不一定能够找到最优模型，而且它可缩放性（scalability）差，当症状变量的数目大时，十分费时. 因此，有必要对 HLC 模型进行更深入的研究，设计出既能够找到最优模型又具良好可缩放性的算法. 就模型的种类而言，除了 HLC 模型以外，还有必要考虑其它类型的隐结构模型. 在 HLC 模型中，一个显变量只能与一个隐变量相连. 在中医应用中，这意味着一个症状的出现只可能由于一种原因. 这个限制与中医实际情况显然不符. 事实上，在对肾虚辨证做研究时所获得的 HLC 模型与肾虚辨证理论有不吻和的地方，其主要原因就是 HLC 模型的这个限制. 所以，有必要取消这个限制，考虑一类比 HLC 模型更为一般的隐结构模型. 另外，症状变量的 4 个取值顺序之间的特殊等级关系在 HLC 模型中被完全忽视，有必要让它在模型中反映出来.

参 考 文 献

董昌武，瘳圣宝，高尔鑫. 2000. 关于中医“证”研究的思考. 中国医药学报，15 (3)，9～11

李小兵，方永奇，洪永敦等. 2000. 心脑血管病痰证宏观辨证的计量化研究. 中国中医基础医学杂志，6 (5)，44～47

梁茂新，刘进，洪志平. 1998. 中医证研究的困惑与对策. 北京：人民卫生出版社

刘士敬，朱倩. 1998. 中医产科病证明脾气虚证型量化论断标准研究——92 例本证型论断因素的多元回归分析. 中国中医基础医学杂志，4 (9)，49～51

陆金宝，周如倩，刘仁人等. 2002. 老年人肾虚及其证型的调查研究. 上海中医药大学学报，16 (1)，22～23

王洪琦，许有玲. 1999. 面向 21 世纪——中医基础理论研究状和未来. 北京：军事医学科学出版社

严石林，张连文，王米渠等. 2001. 肾虚证辨证因子等级评判操作标准的研究. 成都中医药大学学报，24 (1)，56～59

杨维益，陈家旭，王天芳等. 1996. 关于中医“证”研究的思考. 中国医药学报，11 (1)，4～6

杨学鹏. 1993. 阴阳——气与变量. 北京：科学出版社

张连文，袁世宏. 2004. 隐结构模型与中医辨证. Technical Report HKUST-CS04-12，Department of Computer Science，The Hong Kong University of Science and Technology

赵国求. 1999. 中医证实质研究释疑与解惑. 医学与哲学，20 (11)，47～50

朱文锋. 1995. 中医论断学. 上海：上海科学技术出版社

Abramson B. 1991. ARCO1：An application of belief networks to the oil market. In UAI'91：Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence. San Francisco：Morgan Kaufmann Publishers. 1～8

Abramson B. 1994. The design of belief network based systems for price forecasting. Computers and Electrical Engineering，20 (2)：163～180

Abramson B，Finizza A J. 1991. Using belief networks to forecast oil prices. International Journal of Forecasting，7 (3)：299～316

Adams I D，Chan M，Clifford P C，et al. 1986. Computer aided diagnosis of acute abdominal pain：a multicentre study. British Medical Journal，293 (6550)：800～804

Akaike H. 1973. Information theory as an extension of the maximum likelihood principle. In V Petrov，F Csaki，editors. Second International Symposium on Information Theory，Budapest：Akailseoniai Kiudo. 267～281

Andreassen S，Jensen F V，Andersen S K，et al. 1989. MUNIN an expert EMG assistant. In J E Desmedt，editor. Computer Aided Electromyography and Expert Systems. Elsevier Science Publishers. 255～277

Arnborg S，Corneil D G，Proskurowski A. 1987. Complexity of finding embedding in a k tree. Journal of SIAM，Algebraic Discrete Methods，8 (2)：277～284

Atte H. 2001. Reliability estimation of safety critical software based systems using Bayesian networks. Technical Report STUK YTO TR198，Helsinki

Bacon P J，Cain J D，Howard D C. 2002. Belief network models of land manager decisions and land use change. Journal of Environmental Management，65 (1)：1～23

Bartholomew D, Knott M. 1999. Latent variable models and factor analysis. New York: Oxford University Press

Batchelor C, Cain J. 1999. Application of belief networks to water management studies. Agricultural Water Management, 40 (1): 51～57

Bauer E, Koller D, Singer Y. 1997. Update rules for parameter estimation in Bayesian networks. In D Geiger, P P Shenoy, editors. Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI 97). San Francisco: Morgan Kaufmann Publishers. 3～13

Baum L E, Petrie T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Statistics, 37: 1554～1563

Becker A, Geiger D. 2001. A sufficiently fast algorithm for finding close to optimal clique trees. Artificial Intelligence, 125 (1-2): 3～17

Beeri C, Fagin R, Maier D, et al. 1983. On the desirability of acyclic database management. Journal of the ACM, 30 (3): 479～513

Beinlich I, Suermondt H, Chavez R, et al. 1989. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In Proceedings of the Second European Conference on Artificial Intelligence in Medicine, 247～256. London: Springer Verlag

Berrou C, Glavieux A, Thitimajshima P. 1993. Near Shannon limit error correcting coding and decoding: Turbo codes. In IEEE International Conference on Communications '93, 2, Geneva. 1064～1070

Bertele U, Brioschi F. 1972. Nonserial dynamic programming. New York: Academic Press

Birney E. 2001. Hidden Markov models in biological sequence analysis. IBM Journal of Research and Development, 45 (3-4): 449～454

Blei D M, Jordan M I, Ng A Y. 2003. Hierarchical Bayesian models for applications in information retrieval. In J M Bernardo, M J Bayarri, A P Dawid, J O Berger, D Heckerman, A F M Smith, M West, editors. Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting. Oxford: Oxford University Press. 25～44

Bockhorst J, Craven M, Page D, et al. 2003. A Bayesian network approach to operon prediction. Bioinformatics, 19 (10): 1227～1235

Bollen K A. 1989. Structural equations with latent variables. New York: John Wiley & Sons Inc. Publishers

Borsboom D, Mellenbergh G J, van Heerden J. 2003. The theoretical status of latent variables. Psychological Review, 110 (2): 203～219

Borsuk M E, Stow C, Reckhow K. 2002. Integrative environmental prediction using Bayesian networks: A synthesis of models describing estuarine eutrophication. In Proceedings of IEMSS Conference '02, part II: Techniques and Methodologies, Lugano, Switzerland. 102～107

Boutilier C, Friedman N, Goldszmidt M, et al. 1996. Context specific independence in Bayesian networks. In F J E H Eric Horvitz Finn Jensen Eric Horvitz, F Jensen, editors. UAI'96: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 115 ～123

Breese J S, Blake R. 1995. Automating computer bottleneck detection with belief nets. In S Besnard, S Hanks, editors. UAI'95: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 36～45

Burnell L, Horvitz E. 1995. Structure and chance: melding logic and probability for software debugging.

Communications of the ACM, 38 (3): 31～41, 57

Burnham K P, Anderson D R. 2002. Model selection and multimodel inference: a practical information theoretic approach. New York: Springer Verlag

Cain J D, Jinapala K, Makin I W, et al. 2003. Participatory decision support for agricultural management. a case study from Sri Lanka. Agricultural Systems, 76 (2): 457～482

Cannings C. 1978. Probability functions on complex pedigrees. Advances in Applied Probability, 10: 26～61

Carnap R. 1950. Logical Foundations of Probability. Chicago: University of Chicago Press

Castelo R, Kocka T. 2003. On inclusion driven learning of Bayesian networks. Journal of Machine Learning Research, 4 (9): 527～574

Cheeseman P, Stutz J. 1995. Bayesian classification (autoclass): Theory and results. In P S UM Fayyad G Piatetsky Shapiro, R Uthurusamy, editors. Advances in Knowledge Discovery and Data Mining. Menlo Park: The AAAI Press. 153～180

Chen T, Kocka T, Zhang N L. 2005. Effective dimensions of partially observed polytrees. International Journal of Approximate Reasoning, 38 (3): 311～332

Cheng J. Druzdzel M J. 2000. AIS BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. Journal of Artificial Intelligence Research, 13: 155～188

Cheng J, Greiner R. 1999. Comparing Bayesian network classifiers. In UAI'99: Proceedings of the fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 101～108

Chickering D M. 2002. Learning equivalence classes of Bayesian network structures. Journal of Machine Learning Research, 2 (3): 445～498

Chickering D M. 2003. Optimal structure identification with greedy search. Journal of Machine Learning Research, 3 (3): 507～554

Chickering D M, Heckerman D. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Machine Learning, 29 (23): 181～212

Cooper G F. 1990. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence, 42 (2-3): 393～405

Cooper G F, Herskovits E. 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9 (4): 309～347

Corney D. 2000. Designing food with Bayesian belief networks. In I Parmee, editor. ACDM 2000: the Fourth International Conference on Adaptive Computing in Design and Manufacture. University of Plymouth, UK. 83～94

Cowell R G, Dawid A P, Lauritzen S L, et al. 1999. Probabilistic Networks and Expert Systems. New York: Springer Verlag

D'Ambrosio B. 1991. Local expression languages for probabilistic dependence: a preliminary report. In UAI '91: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 95～102

D'Ambrosio B. 1993. Incremental probabilistic inference. In D Heckerman, E H Mamdani, editors. UAI'93: Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 301～308

D'Ambrosio B, Burgess S. 1996. Some experiments with real time decision algorithms. In E Horvitz, F Jensen, editors. UAI'96: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 194～202

Dawid A P. 1992. Applications of a general propagation algorithm in a probabilistic expert system. Statistics and Computing, 2: 25～36

Dawid A P. 2003. An object oriented Bayesian network for estimating mutation rates. In C M Bishop, B J Frey, editors. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Key West. 3～6

Dawid A P, Mortera J, Pascali V L, et al. 2002. Probabilistic expert systems for forensic inference from genetic markers. Scandinavian Journal of Statistics, 29 (4): 577～595

de Dembal F T, Leaper D J, Staniland J R, et al. 1972. Computer aided diagnosis of acute abdominal pain. British Medical Journal, 2 (5804): 9～13

Dean T, Kanazawa K. 1990. A model for reasoning about persistence and causation. Computational Intelligence, 5 (3): 142～150

Dechter R. 1999. Bucket elimination: A unifying framework for reasoning. Artificial Intelligence, 113 (1-2): 41～85

Dempster A P, Laird N M, Rubin D B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal statistical Society, 39 (1): 1～38

Deng J. 1999. Researches on some key problems of Turbo codes [Ph. D. thesis], Shanghai University

Dittmer S L, Jensen F V. 1997. Tools for explanation in Bayesian networks with application to an agricultural problem. In Proceedings of the First European Conference for Information Technology in Agriculture, Copenhagen, Denmark. 15～18

Draper D. 1995. Localized Partial Evaluation of Bayesian Belief Networks [Ph. D. thesis], University of Washington

Druzdzel M J. 1994. Some properties of joint probability distributions. In D Poole and Ramon Lopez de Mantaras, D Poole, editors. UAI'94: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 187～194

Duda R O, Hart P E. 1973. Pattern classification and scene analysis. New York: John Wiley & Sons Inc. Publishers

Duda R O, Hart P E, Neisson N J. 1976. Subjective Bayesian methods for rule based inference systems. In Proceedings of the 1976 National Computer Conference, volume 45. AFIPA Press. 1075～1082

Duda R O, Hart P E, Stack D G. 2001. Pattern classification. New York: John Wiley & Sons Inc. Publishers

Durbin R, Eddy S, Krough A, et al. 1988. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press

Edwards D. 2000. Introduction to graphical modelling. New York: Springer Verlag

Elidan G, Friedman N. 2001. Learning the dimensionality of hidden variables. In J Breese, D Koller, editors. UAI'01: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,. 144～151

Elidan G, Lotner N, Friedman N, et al. 2000. Discovering hidden variables: A structure based approach. In Proceedings of the 13th Annual Conference on Neural Information Processing Systems, Denver, Colorado.

479～485

Everitt B S, Hand D. 1981. Finite Mixture Distributions. London: Chapman and Hall

Fishelson M, Geiger D. 2003. Optimizing exact genetic linkage computations. In The Seventh Annual International Conference on Research in Computational Molecular Biology, Berlin. 114～121

Freeman W T, Pasztpr E C, Carmichael O T. 2000. Learning low level vision. International Journal of Computer Vision, 40 (1): 25～47

Frey B J. 1998. Graphical Models for Machine Learning and Digital Communication. Cambridge: The MIT Press

Frey B J, Mackay D J C. 1998. A revolution: belief propagation in graphs with cycles. In Advances in Neural Information Processing Systems, volume 10. Cambridge: The MIT Press

Friedman N, Geiger D, Goldszmidt M. 1997. Bayesian network classifiers. Machine Learning, 29 (2-3): 131～163

Friedman N, Geiger D, Lotner N. 2000a. Likelihood computation with value abstraction. In C Boutilier, M Goldszmidt, editors. UAI'00: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 192～200

Friedman N, Linial M, Nachman I, et al. 2000b. Using Bayesian networks to analyze expression data. Journal of Computational Biology, 7 (3-4): 601～620

Friedman N, Ninio M, Pe'er I, et al. 2002. A structural EM algorithm for phylogenetic inference. Journal of Computational Biology, 9: 331～353

Friis Hansen A. 2001. Bayesian networks as a decision support tool in marine applications [Ph. D. thesis]. Lyngby, Denmark

Fung R, Chang K C. 1989. Weighting and integrating evidence for stochastic simulation in Bayesian networks. In M Henrion, R D Shachter, L N Kanal, J F Lemmer, editors. UAI'89: Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence, North-Holland. 209～220

Fung R, Favero B D. 1994. Backward simulation in Bayesian networks. In R L de Mantaras, D Poole, editors. UAI'94: Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 227～234

Fung R, Favero B D. 1995. Applying Bayesian networks to information retrieval. Communications of the ACM, 48 (3): 42～48

Gebhatdr J, Detmer H, Madsen A L. 2003. Predicting parts demand in the automotive industry: An application of probabilistic graphical models. In UAI'03: Bayesian Modeling Application Workshop. San Francisco: Morgan Kaufmann Publishers

Geiger D, Heckerman D, et al. 1996. Asymptotic model selection for directed networks with hidden variables. In UAI'96: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 158～168

Ghahramani Z. 2001. An introduction to hidden Markov models and Bayesian networks. International Journal of Pattern Recognition and Artificial Intelligence, 15 (1): 9～42

Gibbs W. 1902. Elementary principles of statistical mechanics. New Haven: Yale University Press

Ginsberg M. 1987. Readings in nonmonotonic reasoning. San Francisco: Morgan Kaufmann Publishers

Glover F. 1986. Future paths for integer programming and links to artificial intelligence. Computers and Operations Research, 13 (5): 533～549

Glover F. 1989. Tabu search part I. ORSA Journal on Computing, 1 (3): 190～206

Goldberg D E. 1989. Genetic algorithms in search, optimization and machine learning. Boston: Addison Wesley

Golumbic M C. 1980. Algorithmic graph theory and perfect graphs. New York: Academic Press

Good I J. 1961. A causal calculus. British Journal of the Philosophy of Science, 11: 305～318

Goodman L A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61: 215～231

Green P J. 1998. Penalized likelihood. In Encyclopaedia of Statistical Sciences, volume 3. New York: John Wiley & Sons Inc. Publishers. 578～586

Grois E, Hsu W H, Voloshin M, et al. 1998. Bayesian network models for automatic generation of crisis management training scenarios. In Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence. 1113～1120

Guo H P. 2003. Algorithm selection for sorting and probabilistic reasoning: a learning based approach [Ph. D. thesis]. Kansas State University

Guo H P, Boddhireddy P, Hsu W. 2004. An ACO algorithm for the MPE problem. In Lecture Notes in Computer Science, volume 3339. Berlin: Springer Verlag. 778～790

Haas T C. 1991. Partial validation of Bayesian belief network advisory systems. AI Applications, 5 (4): 59～71

Haas T C. 1992. A Bayes network model of district ranger decision making. AI Applications, 6 (3): 72～88

Haas T C, Mowrer H T, Shepperd W D. 1994. Modeling aspen stand growth with a temporal Bayes network. AI Applications, 8 (1): 15～28

Hagnaars J Q, Mccutcheon A. 2002. Applied latent class analysis models. Cambridge: Cambridge University Press

Hautaniemi S, Edgren H, Vesanen P, et al. 2003. A novel strategy for microarray quality control using Bayesian networks. Bioinformatics, 19 (16): 2031～2038

Hautaniemi S, Korpisaari P, Saarinen J. 2000. Target identification with dynamic hybrid Bayesian networks. In Proceedings of the EOS/SPIE Symposium on Remote Sensing, Barcelona, Spain. 92～102

Hautaniemi S, Saarinen J. 2001. Multitarget tracking with IMM and Bayesian networks: Empirical studies. In B V Dasarathy, editor. Proceedings of the SPIE International Symposium on Sensor Fusion: Architectures, Algorithms, and Applications V, volume 4385, Orlando, FL. 47～57

Heckerman D. 1985. Probabilistic interpretation for MYCIN's certainty factors. In L N Kanal, J F Lemmer, editors. UAI'85: Proceedings of the First Conference on Uncertainty in Artificial Intelligence. New York: Elsevier. 167～196

Heckerman D. 1988. An empirical comparison of three inference methods. In UAI'88: Proceedings of the 4th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 158～169

Heckerman D. 1990. Probabilistic similarity networks. Networks, 20: 607～636

Heckerman D. 1993. Causal independence for knowledge acquisition and inference. In UAI'93: proceedings of the 9th Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 122～127

Heckerman D, Breese J, Rommelse K. 1995. Decision theoretic troubleshooting. Communications of the

ACM, 38 (3): 49～57

Henrion M. 1988. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In J Lemmer, L Kanal, editors. UAI'88: Proceedings of the second Conference on Uncertainty in Artificial Intelligence, North-Holland. 149～163

Hojsgaard S, Rasmussen H H, Djurhuus J. 2002. Erospredict - a program for predicting soil erosion. In Proc. 4th EFITA 2003, Debrecen-Budapest, Hungary. 1～8

Holland J H. 1975. Adaptation in natural and artificial systems. Ann Arbor: The University of Michigan Press

Horvitz E. 1990. Computation and Action Under Bounded Resources [Ph. D. thesis]. Stanford University

Horvitz E, Barry B. 1995. UAI'95: Display of information for time critical decision making. In P Besnard, S Hanks, editors. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 296～305

Horvitz E, Breese J, Heckerman D, et al. 1998. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In UAI'98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 256～265

Howard R, Matheson J E. 1984. Influence diagrams. In R Howard, J E Matheson, editors. Readings on the Principles and Applications of Decision Analysis, volume II. Strategic Decisions Group, Menlo Park, CA. 721～762

Jaakkola T M, Jordan M I. 1999. Variational probabilistic inference and the QMRDT network. Journal of Artificial Intelligence Research, 10: 291～322

Jaakkola T S, Jordan M I. 1996. Computing upper and lower bounds on likelihoods in intractable networks. In UAI'96: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence. 340～348

Jensen F, Andersen S. 1990. Approximations in Bayesian belief universes for knowledge based systems. In UAI'90: Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence. 162～169

Jensen F V. 1996. An Introduction to Bayesian Networks. New York: Springer Verlag

Jensen F V. 2001. Bayesian networks and decision networks. New York: Springer Verlag

Jensen F V, Kjaerulff U, Kristiansen B, et al. 2001. The SACSO methodology for troubleshooting complex systems. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 15: 321～333

Jensen F V, Olesen K G, Anderson K. 1990. An algebra of Bayesian belief universes for knowledge based systems. Networks, 20: 637～659

Jitnah N, Nicholson A E. 1996. Belief network algorithms: A study of performance based on domain characterization. In Lecture Notes In Computer Science: Selected Papers from the PRICAI Workshop, volume 1359. London: Springer Verlag. 168～187

Jolliffe I. 1988. Principal component analysis. New York: Springer Verlag

Jordan M, editor. 1998. Learning in graphical models. Cambridge: MIT Press

Jordan M. 2004. Graphical models. Statistical Science (Special Issue on Bayesian Statistics), 19: 140～155

Jordan M, Ghahramani Z, Jaakkola T, et al. 1999. An introduction to variational methods for graphical models. Machine Learning, 37 (2): 183～233

Jorgensen E, Toft N. 1999. A Bayesian network based monitoring system for sow management. Technical Report DINA Research Report 80

Kalman R E. 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME

Journal of Basic Engineering, 82 (Series D): 35～45

Kass R, Tierney L, Kadane J. 1988. Asymptotic in Bayesian computation. In J Bernardo, M DeGroot, D Lindley, A Smith, editors. Bayesian Statistics 3. Oxford: Oxford University Press. 261～278

Kass R E, Vos P. 1997. Geometrical Foundations of Asymptotic Inference. New York: John Wiley & Sons Inc. Publishers

Kemmerer B, Mishra S, Shenoy P P. 2002. Bayesian causal maps as decision aids in venture capital decision making: Methods and applications. In Proceedings of the Academy of Management Conference

Kim J H, Pearl J. 1983. A computational model for combined causal and diagnostic reasoning in inference systems. In IJCAI'83, Karsruhe, Germany. 190～193

Kindermann R, Snell J L. 1980. Markov random fields and their applications. AMS

Kipersztok O, Provan G. 2003. A framework for diagnostic inference of commercial aircraft systems. In UAI03 Bayesian Modeling Applications Workshop. Acapulco

Kirkpatrick S, Gelatt C D, Vecchi M P. 1983. Optimization by simulated annealing. Science, 220: 671～680

Kjaerulff U. 1994. Reduction of computation complexity in Bayesian networks through removal of weak dependencies. In UAI'94: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 374～382

Kocka T, Zhang N L. 2002. Dimension correction for hierarchical latent class models. In UAI'02: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. 267～274

Kolmogorov A N. 1933. Grundbegriffe der Wahrscheinlichkeitsrechnung (Foundations of Probability). Berlin: Springer Verlag

Kolmogorov A N. 1950. Foundations of probability. New York: Chelsea

Kristensen K, Rasmussen I A. 2002. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. Computers and Electronics in Agriculture, 33 (3): 197～217

Kschischang F R, Frey B J. 1998. Iterative decoding of compound codes by probability propagation in graphical models. IEEE Journal on Selected Areas in Communications, 16: 219～230

Kschischang F R, Frey B J, Loeliger H A. 2001. Factor graphs and the sum product algorithm. IEEE Transactions on Information Theory, 47: 498～519

Langley P, Iba W, Thompson K. 1992. An analysis of Bayesian classifiers. In AAAI'92: Proceedings of the Tenth National Conference on Artificial Intelligence. 223～228

Langseth H. 2002. Bayesian networks with applications in reliability analysis [Ph. D. thesis]. Norwegian University of Science and Technology

Lanterman A D. 2001. Schwarz, Wallace, and Rissanen: Intertwining themes in theories of model selection. International Statistical Review, 69: 185～212

Laskey K, Laskey G. 2002. Combat identification with Bayesian networks. In Command and Control Research and Technology Symposium 2002

Lauritzen S. 1996. Graphical models. Oxford: Clarendon Press

Lauritzen S L. 1995. The EM algorithm for graphical association models with missing data. Computational Statistics and Data Analysis, 19 (2): 191～201

Lauritzen S L, Jensen F V. 1997. Local computation with valuations from a commutative semi group. An-

nals of Mathematics and Artificial Intelligence, 21: 51～69

Lauritzen S L, Spiegelhalter D J. 1988. Local computations with probabilities on graphical structures and their application to expert system. Royal Statistic Society, 50: 154～227

Lazarsfeld P F, Henry N W. 1968. Latent structure analysis. Boston: Houghton Mifflin

Lee D C, Rieman B E. 1997. Population viability assessment of salmonids by using probabilistic networks. North American Journal of Fisheries Management, 17: 1144～1157

Lerner U, Parr R, Koller D, et al. 2000. Bayesian fault detection and diagnosis in dynamic systems. In AAAI/IAAI 2000. 531～537

Linhart H, Zucchini W. 1986. Model selection. New York: John Wiley & Sons Inc. Publishers

Littman M, Majercik S, Pitassi T. 2001. Stochastic Boolean satisfiability. Journal of Automated Reasoning, 27 (3): 251～296

Liu E, Zhang D. 2003. Diagnosis of component failures in the space shuttle maine engines using Bayesian belief networks: A feasibility study. International Journal on Artificial Intelligence Tools, 12: 355～374

Mackay D J C. 2004. Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press

Madsen A L, Jensen F V. 1998. Lazy propagation in junction trees. In G F Cooper, S Moral, editors. UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 362～369

Maier D. 1983. The theory of relational databases. Rockville: Computer Science Press

Marcot B G, Holthausen R S, Raphael M G, et al. 2001. Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. Forest Ecology and Management, 153 (1-3): 29～42

Markov A A. 1912. Wahrscheinlichkeitsrechnung (German translation of 2nd Russian edition: Markoff, A. A., 1908). Leipzig: Teubner

McCarthy J. 1980. Circumscription: A form of non monotonic reasoning. Artificial Intelligence, 13: 27～39

McDermott D, Doyle J. 1987. Non monotonic logic I. In M L Ginsberg, editor. Readings in Nonmonotonic Reasoning. Los Altos: Kaufmann. 111～126

McEliece R J, MacKay D J C, Cheng J F. 1998. Turbo decoding as an instance of Pearl's belief propagation algorithm. Journal on Selected Areas in Communications, 16 (2): 140～151

McLachlan G, Krishnan T. 1997. The EM algorithm and extensions. New York: John Wiley & Sons Inc. Publishers

McLachlan G J, Basford K. 1988. Mixture models: Inference and applications to clustering. New York: Marcel Dekker

Meila M, Jordan M. 1997a. Triangulation by continuous embedding. In M C Mozer, M Jordan, T Petsche, editors. Advances in Neural Information Processing Systems 9. Cambridge: MIT Press

Meila M, Jordan M I. 1997b. An objective function for belief net triangulation. In D Madigan, P Smyth, editors. Proceedings of the 1997 Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL

Mengshoel O J, Wilkins D C. 1997. Visualizing uncertainty in battlefield reasoning using belief networks. In Proceedings of the ARL Advanced Displays and Interactive Displays Federated Laboratory First Annual Symposium. Adelphi, MD, ARL Adelphi Laboratory. 15～22

Mengshoel O J, Wilkins D C, Uckun S. 1998. Filtering and visualizing uncertain battlefield data using

Bayesian networks. In Proceedings of the ARL Advanced Displays and Interactive Displays Federated Laboratory Second Annual Symposium. College Park, MD

Metropolis N, Rosenbluth A, Rosenbluth M, et al. 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics, 21: 1087～1091

Miller R A, Masarie F E, Myers J D. 1986. Quick medical reference (QMR) for diagnostic assistance. Medical Computing, 3: 34～48

Miller R A, Pople H E J, Myers J D. 1982. Internist 1: An experimental computer based diagnostic consultant for general internal medicine. New England Journal of medicine, 307 (8): 468～476

Minka T. 1999. From hidden Markov models to linear dynamic systems. Technical Report TR (531. Cambridge: MIT Press

Moore R C. 1987. Semantical considerations on nonmonotonic logic. Artificial Intelligence, 25: 75～94

Morjaia M, Rink J, Smith W, et al. 1993. Commercialization of EPRI's generator expert monitoring system (gems). In Expert System Application for the Electric Power Industry

Morris A T, Beling P A. 2001. Space shuttle RTOS Bayesian network. In Proceedings of the 20th Digital Avionics Systems Conference. Daytona Beach, Florida. 2001

Murphy K, Weiss Y, Jordan M. 1999. Loopy belief propagation for approximate inference: An empirical study. In UAI'99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 467～475

Murphy K P. 2002. Dynamic Bayesian Networks: Representation, Inference and Learning [Ph. D. thesis]. University of California, Berkeley

Musman S A, Booker L B. 1999. Real time scheduling under uncertainty for ship self defense. Submitted to IEEE Expert, Special Issue on Real Time Intelligent System

Musman S A, Chang L W, Booker L B. 1990. A real time control strategy for Bayesian belief networks with application to ship classification problem solving. In Tools for AI Conference. Washington, D. C.

Musman S A, Chang L W, Booker L B. 1993. Application of a real time control strategy for Bayesian belief networks to ship classification. International Journal for Pattern Recognition and Artificial Intelligence, 7 (3): 513～526

Neal R M. 1993. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG TR-93-1, University of Toronto

Neapolitan R E. 2003. Learning Bayesian networks. New York: Prentice Hall

Neil M, Tranham E. 2002. Using Bayesian networks to predict OpRisk. Operational Risk, 8～10

Nilsson N J. 1965. Learning machines. New York: McGraw Hill

Park J. 2002. MAP complexity results and approximation methods. In UAI'02: Proceedings of the 18th Annual Conference on Uncertainty in AI. San Francisco: Morgan Kaufmann Publishers. 388～396

Parker R C, Miller R A. 1987. Using causal knowledge to create simulated patient cases: the CPCS project as an extension of INTERNIST-1. In Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. IEEE Comp Soc Press. 473～480

Pearl J. 1982. Reverend Bayes on inference engines: A distributed hierarchical approach. In AAAI-82. Pittsburgh. 133～136

Pearl J. 1986. Fusion, propagation, and structuring in belief networks. Artificial Intelligence, 29 (3): 241～288

Pearl J. 1987. Evidential reasoning using stochastic simulation of causal models. Artificial Intelligence, 32 (2): 245～257

Pearl J. 1988. Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Morgan Kaufmann

Pearl J. 2000. Causality. Cambridge: Cambridge University Press

Peterson C, Anderson J R. 1987. A mean field theory learning algorithm for neural networks. Complex Systems, 1: 995～1019

Poole D. 1993. The use of conflicts in searching Bayesian networks. In UAI'93: Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 359～367

Popper K. 1957. The propensity interpretation of the calculus of probability and the quantum theory. In S Körner, editor. The Colston Papers, 9, 65～70

Portinale L, Bobbio A. 1999. Bayesian networks for dependability analysis: an application to digital control reliability. In UAI' 99: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. 551 ～558

Rabiner L, Juang B H. 1993. Fundamentals of speech recognition. Englewood Cliffs: Prentice Hall

Rabiner L R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In Proceedings of the IEEE, volume 77. 257～286

Raftery A. 1995. Bayesian model selection in social research. Sociological Methodology, 25: 111～196 (With discussion by Andrew Gelman, Donald B. Rubin, and Robert M. Hauser, and a rejoinder)

Reiter R. 1987. A logic for default reasoning. In M L Ginsberg, editor. Readings in Nonmonotonic Reasoning. Los Altos: Kaufmann. 68～93

Rieman B, Peterson J T, Clayton J, et al. 2001. Evaluation of potential effects of federal land management alternatives on trends of salmonids and their habitats in the interior Columbia river basin. Forest Ecology and Management, 153 (13): 43～62

Rissanen J. 1978. Modeling by shortest data description. Automatica, 14: 465～471

Robinson R W. 1977. Counting unlabeled acyclic digraphs. In C H C Little, editor. Combinatorial Mathematics, Lecture Notes in Mathematics, volume 622. Berlin: Springer Verlag. 28～43

Rose D. 1970. Triangulated graphs and the elimination process. J. Math. Anal. Appl. , 32: 579～609

Roth D. 1996. On the hardness of approximate reasoning. Artificial Intelligence, 82 (1-2): 273～302

Roweis S, Ghahramani Z. 1999. A unifying review of linear Gaussian models. Neural Computation, 11 (2): 305～345

Rubin D R. 1976. Inference and missing data. Biometrica, 63: 581～592

Rubinstein R Y. 1981. Simulation and the Monte Carlo method. New York: John Wiley & Sons Inc. Publishers

Ruokangas C C, Mengshoel O J. 2003. Information filtering using Bayesian networks: effective user interfaces for aviation weather data. In Proceedings of the 2003 International Conference on Intelligent User Interfaces. New York: ACM Press

Russell S, Norvig P. 2003. Artificial intelligence: A modern approach (2nd Edition). NJ: Prentice Hall

Russell S J, Binder J, Koller D, et al. 1995. Local learning in probabilistic networks with hidden variables. In IJCAI-95. 1146～1152

Sahami M, Dumais S, Heckerman D, et al. 1998. A Bayesian approach to filtering junk E-mail. In Learning

for Text Categorization: Papers from the 1998 Workshop. Madison, Wisconsin: AAAI Technical Report WS 98-05

Said A, Stevens D K, Sehlke G. 2001. Exploration of conservations schemes and TMDL using Bayesian networks. In 2001 INRA and INEEL Subsurface Symposium. Idaho Falls

Santos E, Shimony S E, Solomon E, et al. 1996. Sample and accumulate algorithms for belief updating in Bayes networks. In UAI'96: Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 477～484

Santos E, Shimony S E, Williams E. 1995. On a distributed anytime architecture for probabilistic reasoning. Technical Report AFIT/EN/TR94-06, Department of Electrical and Computer Engineering, Air Force Institute of Technology

Santos E J, Shimony S E. 1998. Deterministic approximation of marginal probabilities in Bayes nets. IEEE Transactions on Systems, Man, and Cybernetics, 28 (4): 377～393

Santoso N I, Darken C, Povh G, et al. 1999. Nuclear plant fault diagnosis using probabilistic reasoning. In Proceedings of the 1999 IEEE Power Engineering Society Summer Meeting, volume 2. 714

Sarkar S. 1993. Using tree decomposable structures to approximate belief networks. In UAI'93: Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence. Washington DC: Morgan Kaufmann

Saul L K, Jaakkola T, Jordan M. 1996. Mean field theory for sigmoid belief networks. Journal of Artificial Intelligence Research, 4: 61～76

Saul L K, Jordan M I. 2000. Attractor dynamics in feed forward neural networks. Neural Computation, 12 (6): 1313～1335

Schiaffino S N, Amandi A. 2000. User profiling with case based reasoning and Bayesian networks. In IBERAMIA SBIA 2000 Open Discussion Track. 12～21

Schwarz G. 1978. Estimating the dimension of a model. Annals of Statistics, 6: 461～464

Shachter R. 1986. Evaluating influence diagram. Operations Research, 34: 871～882

Shachter R D, D'Ambrosio B, Favero B D. 1994. Symbolic probabilistic inference in belief networks. In Proceedings of the Eighth National Conference on Artificial Intelligence. 126～131

Shachter R D, Peot M A. 1989. Simulation approaches to general probabilistic inference on belief networks. In UAI'89: Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence. New York: Elsevier Science Publishing Company. 221～231

Shafer G. 1996. Probabilistic expert system. Philadelphia: Society for Industrial and Applied Mathematics

Shafer G, Pearl J. 1990. Readings in uncertain reasoning. San Francisco: Morgan Kaufman Publishers

Shafer G, Shenoy P. 1990. Probability propagation. Annals of Mathematics and Artificial Intelligence, 2: 327～352

Shafter G. 1976. A Mathematical theory of evidence. Princeton: Princeton University Press

Shalom Y B, Li Y. 1993. Estimation and Tracking: Principles, Techniques and Software. Artech House

Shannon C E. 1948. A mathematical theory of communication. Bell System Technical Journal, 27: 379～423

Shenoy C, Shenoy P P. 1999. Bayesian network models of portfolio risk and return. In Y S Abu Mostafa, B LeBaron, A W Lo, A S Weigend, editors. Computational Finance. 87～106

Shenoy P. 1992. Valuation based systems for Bayesian decision analysis. Operations Research, 40 (3): 463～484

Shimony S E. 1994. Finding MAPs for belief networks is NP hard. Artificial Intelligence, 68 (2): 399～410

Shortliffe E H. 1976. Computer based medical consultations: MYCIN. New York: Elsevier

Shortliffe E H, Buchanan B G. 1975. A model of inexact reasoning in medicine. Mathematical Bioscience, 23: 351～379

Shwe M A, Middleton B, Heckerman D E, et al. 1991. Probabilistic diagnosis using a reformulation of the INTERNIST 1/QMR knowledge base: I. the probabilistic model and inference algorithms. Methods of Information in Medicine, 30 (4): 241～255

Silva R, Scheines R, Glymour C, et al. 2003. Learning measurement models for unobserved variables. In UAI'03: Proceedings of the Nineteenth Conference on Uncertainty on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 543～550

Smith J Q. 1989. Influence diagrams for statistical modeling. Annals of Statistics, 17: 654～672

Sobel M E. 1994. Causal inference in latent variable models. In A von Eye, C Clogg, editors. Latent Variables Analysis: Applications for Developmental Research. Thousand Oaks. 3～35

Spetzler C, von Holstein C S. 1975. Probability encoding in decision analysis. Management Science, 22 (3): 340～358

Spiegelhalter D J, Dawid A P, Lauritzen S L, et al. 1993. Bayesian analysis in expert systems. Statistical Science, 8: 219～247

Spirtes P, Glymour C, Scheines R. 2000a. Causation, prediction, and search. New York: MIT Press

Spirtes P, Glymour C, Scheines R. 2001. Causation, prediction, and search (2nd edition, revised). Cambridge: MIT Press

Spirtes P, Glymour C, Scheines R, et al. 2000b. Constructing Bayesian network models of gene expression networks from microarray data. In Proc. of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology

Spirtes P, Meek C. 1995. Learning Bayesian networks with discrete variables from data. In Proceedings of First International Conference on Knowledge Discovery and Data Mining, Montreal, QU. San Francisco: Morgan Kaufmann Publishers. 294～299

Stone M. 1977. Asymptotics for and against cross validation. Biometrika, 64: 19～35

Strimmer K, Moulton V. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol., 17: 875～888

Tarjan R, Lueker G, Rose D. 1976. Algorithmic aspects of vertex elimination on graphs. SIAM Journal on Computing, 5: 266～283

Tarjan R E, Yannakakis M. 1984. Simple linear time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. SIAM Journal on Computing, 13 (3): 566～579

Titterington D M. 1976. Updating a diagnostic system using unconfirmed cases. Appl. Statist., 25: 238～247

Titterington D M, Murray G D, Murray L S, et al. 1981. Comparison of discrimination techniques applied to a complex data set of head injured patients (with discussion). Journal of the Royal Statistical Society, Series A, 144: 145～175

Torres Toledano J G, Sucar L E. 1998. Bayesian networks for reliability analysis of complex systems. In

Lecture Notes in Artificial Intelligence, volume 1484. 195～206

van Engelen R A. 1997. Approximating Bayesian belief networks by arc removal. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (8): 916～920

Verma T. 1987. Causal networks: Semantics and expressiveness. Technical Report R 65, UCLA Cognitive Systems Laboratory

Verma T, Pearl J. 1988. Causal networks: semantics and expressiveness. In UAI'88: Proceedings of the Fourth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 352～359

Vomlel J. 2003. Two applications of Bayesian networks. In Proceedings of conference Znalosti 2003. Ostrava. 73～82

von Neumann J, Morgenstern O. 1947. Theory of games and economics behavior. Princeton: Princeton University Press

Wainwright M J, Jordan M I. 2003. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley

Warner H R, Toronto A F, Veasey L G, et al. 1961. A mathematical approach to medical diagnosis application to congenital heart disease. Journal of the American Medical Association, 177: 177～184

Weidl G, Madsen A L, Dahlquist E. 2003a. Applications of object oriented Bayesian networks for causal analysis of process disturbances. In proceedings of the 44th Scandinavian Conference on Simulation and Modeling. Vasteras

Weidl G, Madsen A L, Dahlquist E. 2003b. Object oriented Bayesian networks for industrial process operation. In UAI03 Bayesian Modeling Application Workshop. San Francisco: Morgan Kaufmann Publishers

Wellman M P, Liu C L. 1994. State space abstraction for anytime evaluation of probabilistic networks. In UAI'94: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers. 567～574

Wermuth N. 1998. Graphical Markov models. In S Kotz, C Read, D Banks, editors. Encyclopedia os Statistical Sciences. New York: John Wiley & Sons Inc. Publishers. 284～300

Wermuth N, Cox D R. 2001. Graphical models: overview. In P B Balter, N J Smelser, editors. International Encyclopedia of the Social and Behavioral Sciences, volume 9. Amsterdam: Elsevier. 6739～6786

Wolbrecht E, D'Ambrosio B, Paasch R, et al. 2000. Monitoring and diagnosis of a multistage manufacturing process using Bayesian networks. AI EDAM, 14 (1): 53～67

Wong S, Butz C. 2000. A Bayesian approach to user profiling in information retrieval. Technology Letters, 4 (1): 50～56

Wright S. 1921. Correlation and causation. Journal of Agriculture Research, 20: 557～585

Yedidia J, Freeman W, Weiss Y. 2000. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. Technical Report TR-2001-16, MERL

Yedidia J S. 2002. Generalized belief propagation and free energy minimization. In Information Theory Workshop at Mathematical Sciences Research Institute (MSRI)

Yedidia J S, Freeman W T, Weiss Y. 2001. Understanding belief propagation and its generalizations. Technical Report TR2001-22, MERL Research Lab

Yedidia J S, Freeman W T, Weiss Y. 2002. Constructing free energy approximations and generalized belief propagation algorithms. Technical Report TR2001-35, MERL

Yildiz B, Golay M, Maynard K P, et al. 2001. Development of expert system with Bayesian networks for application in nuclear power plants. In 2001 EPRI International Maintenance Conference. Houston, TX

Zadeh L A. 1965. Fuzzy sets. Information and Control, 8: 338～353

Zhang N L. 1993. Studies on hypergraphs I: hyperforests. Discrete Applied Mathematics, 42: 95～112

Zhang N L. 1996. Irrelevance and parameter learning in Bayesian networks. Artificial Intelligence, 88: 359～373

Zhang N L. 1998a. Computational properties of two exact algorithms for Bayesian networks. Applied Intelligence, 9 (2): 173～183

Zhang N L. 1998b. Probabilistic inference in influence diagrams. Computational Intelligence, 14 (4): 475～497

Zhang N L. 2004. Hierarchical latent class models for cluster analysis. Journal of Machine Learning Research, 5 (6): 697～723

Zhang N L, Kocka T. 2004a. Effective dimensions of hierarchical latent class models. Journal of Artificial Intelligence Research, 21: 1～17

Zhang N L, Kocka T. 2004b. Efficient learning of hierarchical latent class models. In Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence. Boca Raton

Zhang N L, Nielsen T D, Jensen F V. 2004. Latent variable discovery in classification models. Artificial Intelligence in Medicine, 30 (3): 283～299

Zhang N L, Poole D. 1994. A simple approach to Bayesian network computations. In Proc. of the Tenth Canadian Conference on Artificial Intelligence. 171～178

Zhang N L, Poole D. 1996. Exploiting causal independence in Bayesian network inference. Journal of Artificial Intelligence Research, 5: 301～328

Zhang N L, Qi R, Poole D. 1994. A computational theory of decision networks. International Journal of Approximate Reasoning, 11 (2): 83～158

Ziv H, Richardson D J. 1997. Constructing Bayesian network models of software testing and maintenance uncertainties. In ICSM'97: Proceedings of the International Conference on Software Maintenance, ICS TR-97-23. Washington: IEEE Computer Society. 100～115

英汉词汇对照

abduction	反绎，82
activated	已被激活的，111
additive white Gaussian noise（AWGN）	加性高斯白噪声，59
Akaike information criterion（AIC）	Akaike 信息准则，181
ancestral set	祖先闭集，71
answer extraction	答案提取，118
artificial intelligence（AI）	人工智能，20
autoepistemic logic	自认知逻辑，20
batch learning	批量学习，150
Bayesian causal networks	贝氏因果网，40
Bayesian estimation	贝叶斯估计，147
Bayesian information criterion（BIC）	贝叶斯信息准则，178
Bayesian MAP estimation	贝叶斯 MAP 估计，148
Bayesian network learning	贝叶斯网学习，144
Bayesian networks	贝叶斯网，34
Bayesian score	贝叶斯评分，176
binary symmetric channel	二值对称信道，59
binomial likelihood function	二项似然函数，146
bits	比特，24
blocked	阻塞，68
cardinality learning	势学习，195
cardinality-hidden variable	势隐变量，194
causal independence	因果机制独立，42
causal Markov assumption	因果马尔可夫假设，41
causal networks	因果网，40
centiMorgan	厘摩，55
chain rule	链规则，15
chance node	机会节点，65
channel coding	信道编码，57
characteristic function	特征函数，153
child	子节点，34

chord	弦，123
chordal graph	弦图，123
circumscription logic	限定逻辑，20
classification	分类，61
clique tree propagation	团树传播法，115
clustering	聚类，61
combat identification	战争身份识别，51
complete data	完整数据组，143
complete sample	完整样本，143
concave function	凹函数，22
conditional distribution	条件分布，13
conditional entropy	条件熵，25
conditional mutual information	条件互信息，30
conditional probability	条件概率，13
conjugate family	共轭分布族，150
consistency	一致性，181
consistent	一致的，157
constraint-based	基于约束的，193
contain	包含，181
context	环境，44
context specific independence	环境独立，44
converging connection	汇连，68
convex function	凸函数，22
counter factual	虚设，40
cover	覆盖，106
cover clique	家族覆盖团，107
cross validation	交叉验证，181
cycle	圈，123
data case	数据例，143
data collection	数据组，143
data completion	数据修补，161
data sample	数据样本，143
decision networks	决策网，65
decision node	决策节点，65
decision theory	决策论，21
deduction	演绎，82
default logic	默认逻辑，20

deficiency	缺边数，95
degree of belief	信度，7
degree of fit	拟合程度，145
dependence map（D-map）	依赖图，78
descendants	后代节点，35
descriptive system	描述型系统，21
diagnostic inference	诊断推理，81
directed acyclic graph	有向无圈图，35
directed cycle	有向圈，35
directed separation（d-separation）	有向分隔（d-分隔），69
Dirichlet family	狄利克雷分布族，151
disease	疾病变量，46
distributionally equivalent	分布等价，181
diverging connection	分连，67
doubly hill climbing	双重爬山法，211
dynamic Bayesian networks（DBN）	动态贝叶斯网，62
edge addition	加边，186
edge deletion	减边，186
edge reversal	转边，186
effective dimension	有效维数，195
elimination	消元，85
elimination cost	消元成本，89
elimination order	消元顺序，88
empirical distribution	经验分布，155
entropy	熵，24
equivalent	等价，182
equivalent sample size	等价样本量，148
eutrophication	藻化现象，52
evidence variables	证据变量，81
expectation maximization（EM）	期望优化，161
expected BIC score	期望 BIC 评分，189
expected loglikelihood function	期望对数似然函数，164
expert system	专家系统，20
explain away	得释，68
explanation	解释，83

factor	因子，85
factor analysis	因子分析，221
factorization	分解，85
findings	症状变量，46
forward sampling	顺序抽样，127
fractional sample	碎权样本，161
fractional updating	碎权更新，168
frequentist	频率学派，21
full Bayesian estimation	完全贝叶斯估计，147
generative distribution	原生分布，155
generative model	原生模型，155
genetic algorithm	遗传算法，188
genetic linkage analysis	基因连锁分析，55
global independence	全局独立，158
global maximum	全局最优，188
graphical models	图模型，65
Hessian matrix	黑塞矩阵，179
hierarchical latent class model（HLCM）	多层隐类模型（HLC 模型），204
hierarchical naïve Bayes model	多层朴素贝叶斯模型，221
holdout validation likelihood	验证数据似然度，181
hyper parameter	超参数，148
hypothesis	假设，82
hypothesis variables	假设变量，82
identifiable	可分辨的，195
imaginary data	虚拟数据，150
importance distribution	重要性分布，125
importance sampling	重要性抽样，125
inactivated	未被激活的，111
incomplete data	缺值数据组，143
incomplete sample	缺值样本，143
independence map（I-map）	独立图，78
independent and identically distributed（i. i. d.）	独立同分布，146
inference	推理，81
information inequality	信息不等式，28

intercausal inference 原因关联推理，81
interior point 内点，178
iterative belief propagation 迭代信度传播，136

joint distribution 联合分布，10
joint entropy 联合熵，25

Laplace approximation 拉普拉斯近似，178
latent class 隐类，197
latent class analysis 隐类分析，200
latent class model (LCM) 隐类模型，197
latent profile analysis 隐概况分析，221
latent structure model 隐结构模型，194
latent trait analysis 隐特质分析，221
latent variable 隐变量，194
latent variable model 隐变量模型，194
leaf node 叶节点，35
likelihood 似然度，145
likelihood function 似然函数，145
linear dynamic system 线性动态系统，64
local independence 局部独立，61
local optimal 局部最优，188
logic sampling 逻辑抽样，127
loglikelihood function 对数似然函数，146
loopy propagation 循环传播法，136

manifest image 显象流形，195
manifest variable 显变量，194
manifestatively contains 显象包含，195
manifestatively equivalent 显象等价，196
manifestatively equivalent in distribution 显象分布等价，195
manifold 流形，195
marginal distribution 边缘分布，11
marginal independent 边缘独立，16
marginal likelihood 边缘似然度，176
marginalization 边缘化，12
Markov boundary 马尔可夫边界，73
Markov chain 马氏链，133

Markov chain Monte Carlo	马尔可夫链蒙特卡洛，125
Markov random field	马尔可夫随机场，206
maximal clique	极大团，123
maximum a posteriori hypothesis (MAP)	最大后验假设，82
maximum cardinality search	最大势搜索，95
maximum clique size	最大团大小，89
maximum expected utility	最大期望效用，21
maximum likelihood estimation (MLE)	最大似然估计，145
mean field equation	平均场方程，135
minimal I-map	最小独立图，78
minimum deficiency search	最小缺边搜索，95
missing at random (MAR)	随机缺失，160
mixed inference	混合推理，82
mixture model	混合模型，221
model	模型，144
model averaging	模型平均，175
model optimization	模型优化，172
model selection	模型选择，172
Monte Carlo algorithms	蒙特卡洛算法，125
moral graph	端正图，74
moralization	端正化，74
most probable explanation (MPE)	最大可能解释，83
multidimensional clustering	多维聚类，211
multinomial likelihood function	多项似然函数，151
mutual information	互信息，27
naïve Bayes classifier	朴素贝叶斯分类器，61
naïve Bayes model	朴素贝叶斯模型，61
naïve mean field theory	朴素平均场理论，134
nats	奈特，24
neighbours	邻居节点，35
node deletion	减节点，212
node introduction	加节点，212
node relocation	移节点，212
noisy adder	噪音加法器，43
noisy AND gate	噪音与门，43
noisy MAX/MIN gate	噪音最大（小）门，43
noisy OR gate	噪音或门，43

non-monotonic inference	非单调推理，20
non-monotonic modal logic	非单调模态逻辑，20
nondescendants	后代节点，35
nonserial dynamic programming (NDP)	不定序动态规划，99
normalization	规一化，110
normalization constant	规一化常数，20
normative system	规范型系统，21
observed variable	观测变量，194
optimal elimination order	最优消元顺序，94
overfitting	过度拟合，174
parallel concatenated convolutional code	并行级联卷积码，59
parameter estimation	参数估计，144
parameter learning	参数学习，144
parameter maximized loglikelihood function	优参似然函数，172
parameter prior	参数先验分布，175
parent	父节点，34
parsimonious model	最简模型，181
penalized likelihood	罚项似然度，181
perfect map (P-map)	完美图，78
phylogenetic tree	进化树，221
plateau	坪区，188
posterior probability	后验概率，18
potential	势位函数，206
pre-HLC model	预 HLC 模型，209
predictive inference	预测推理，81
principal component analysis	主成分分析，221
prior probability	先验概率，18
probabilistic inference	概率推理，81
product Dirichlet distribution	乘积狄利克雷分布，158
production system	产生式系统，20
propensity interpretation	特性解释，9
query	查询，81
query variables	查询变量，81
random interleaver	随机交织器，59

random number generator 随机数产生器，127
random restart hill-climbing 随机重复爬山法，188
random sampling algorithms 随机抽样算法，125
rational agent 理性个体，8
regularity 正则性，198
regularization 正则化，199
rejection sampling 舍选抽样，129
relative entropy 相对熵，28
root node 根节点，35
root walking 走根，205
rule-based expert system 基于规则的专家系统，20

score-based 基于评分的，193
scoring function 评分函数，172
search operator 搜索算子，186
separator 分隔集，119
sequential learning 顺序学习，150
serial connection 顺连，66
simulated annealing 模拟退火，188
stationary distribution 平稳分布，133
strictly concave 严格凹，22
structural graph 结构图，90
structure EM algorithm 结构 EM 算法，188
structure learning 结构学习，144
structured equation model 结构方程模型，221
structure prior 结构先验分布，175
subjective probability 主观概率，7
sufficient statistics 充分统计量，146
super set 超集，123

tabu search 禁忌搜索，188
testing data 测试样本，217
topological sort 拓扑序，127
training data 训练数据，181
tree augmented naïve Bayes model（TAN） 加树朴素贝叶斯模型，62
triangulation 三角化，123

uncertain reasoning 不确定性推理，31

uncertainty 不确定性，21
undirected sepatation（u-separation） 无向分隔（u-分隔），75
unidentifiability of root 根不可分辨性，206
unidentifiable 不可分辨的，195
utility node 效用节点，65

validation data 验证数据，181
value-hidden variable 值隐变量，194
variable elimination（VE） 变量消元法，83
variable-connectedness 变量连通性，106
variational method 变分法，133
variational optimization 变分优化，133
variational transformation 变分转换，133
Venn diagram 韦恩图，27

索　引

A

凹函数，22

B

贝叶斯公式，19
贝叶斯网，34，35
　贝氏因果网，40
　贝叶斯网的构造，36
　贝叶斯网对其它领域的影响，53
　贝叶斯网的应用，45
　动态贝叶斯网，62
　确定网络参数，41
　时变贝叶斯网，63
　因果关系与贝叶斯网，39
　因果网，40
贝叶斯网学习，141，144
　贝叶斯估计，147，148
　参数学习，143，144
　结构学习，144
　批量学习，150
　顺序学习，150
　最大似然估计，145
边缘独立，29
边缘概率分布，11
变量独立，29
变量连通性，106
变量消元 MAP 算法，102
变量消元法，83
标准维数，195
不定序动态规划，99
不可分辨的，195

C

参数学习，143
查询变量，81
超参数，148
乘积狄利克雷分布，158
充分统计量，146

D

单维聚类，211
等价样本量，148
独立，16
　边缘独立，16
　变量独立，16
　局部独立，61
　事件独立，16
　条件独立，16
独立同分布，145
独立图，78
端正图，74
多层隐类模型，204
多维聚类，211
多元概率，10

F

罚项似然度，181
反绎，82
分布等价，181

G

概率的解释，6
　古典解释，6
　逻辑解释，9

频率解释，6
特性解释，9
主观解释，7
概率评估，7
概率推理，81
概率质量函数，5
共轭分布族，150
观测变量，194
规一化常数，20
过度拟合，174
$\mathscr{G}$-可分解的，77

H

黑塞矩阵，179
后验概率，18
后验概率问题，81
互信息，27，29
环境独立，44
结构 EM 算法，188

J

结构学习，172
近似推理，125
变分法，133
基于搜索的算法，138
模型简化法，138
随机抽样算法，125
吉布斯抽样，132
逻辑抽样，127
顺序抽样，127
似然加权法，129
重要性抽样算法，125
MCMC 抽样，131
经验分布，155
局部独立，61，158
局部马尔可夫性，74
局部最优，188
决策论，21
决策网，65
Jensen 不等式，22

K

卡尔曼滤波器，64
Kolmogorov 公理，4
Kullback-Leibler 距离，28

L

拉普拉斯近似，178
联合分布，10
联合概率分布，10
链规则，15

M

马尔可夫边界，73
马尔可夫随机场，206
描述型系统，21
模拟退火，188
模型，144
贝叶斯网模型，144
模型参数，144
模型结构，144
模型平均，175
模型选择，172
贝叶斯模型选择，175
参数先验分布，175
结构先验分布，175
评分函数，172
贝叶斯评分，176
边缘似然函数，176
家族 BIC 评分，183
期望 BIC 评分，189
优参对数似然函数，172
最大优参似然准则，173
AIC 评分，181
BIC 评分，178
CH 评分，178

HVL 评分，181
MDL 评分，180
似然函数，172
最简模型，181
模型优化，182
爬山法，186
评分函数的分解，183
K2 算法，184

N

拟合程度，145

P

频率学派，21
朴素贝叶斯模型，61

Q

期望优化，161
全局独立，158
全局最优，188
缺值数据结构学习，188

S

熵，24
联合熵，25
熵的链规则，27
条件熵，25
相对熵，28
最大熵原理，25
事件，3
数据修补，161
数据样本，143
缺值样本，143
数据例，143
完整样本，143
数据组，143
缺值数据组，143
完整数据组，143
双重爬山法，211
对数似然函数，146
多项似然函数，151
二项似然函数，146
期望对数似然函数，164
似然函数，145
搜索算子，212
随机变量，4
随机缺失，160
随机试验，3
碎权更新，168
碎权样本，161

T

条件独立，30
条件分布，13
凸函数，22
图模型，65
团树，106
团树传播算法，106
TAN 模型，62

U

u-分隔，75

W

完美图，78
完全贝叶斯网，174
无根 HLC 模型，206

X

先验概率，18
显变量，194
显象包含，195
显象等价，196
显象分布等价，195
线性动态系统，64
信息不等式，28
虚拟数据，150

Y

演绎，82
样本点，3
样本空间，3
一致性，181
依赖图，78
已被激活的，111
因果机制独立，42
因果马尔可夫假设，41，74
隐变量，194
　　势隐变量，194
　　值隐变量，194
隐变量模型，194
隐结构法，225，227
隐结构模型，194
隐类模型，197
隐马尔可夫模型，64
有向分隔，74，75
有向无圈图，35
有效维数，195
原生分布，155
原生模型，155

Z

整体马尔可夫性，72
证据变量，81
中医辨证，225
　　回归分析，226
　　聚类分析，227
　　判别分析，226
　　肾虚辨证，229
　　证，225
　　证候，225
　　证实质研究，225
状态空间，4
走根，205
最大后验假设问题（MAP），82
最大可能解释问题（MPE），83
最大期望效用，21
最小独立图，78
最小缺边搜索，95