國 立 交 通 大 學

統 計 學 研 究 所

碩 士 論 文

利用基因網路分析變數間之因果關係

Causal Analysis with Gene Network

研究生：林育仕

指導教授：洪慧念 博士

中華民國 九十四 年 六 月

利用基因網路分析變數間之因果關係

Causal Analysis with Gene Network

研究生：林育仕　　Student：Yu-Shi Lin

指導教授:洪慧念 博士　Advisor:Dr. Hui-Nien Hung

國立交通大學理學院

統計研究所

碩士論文

A thesis

Submitted to Institute of Statistics

College of Science

National Chiao Tung University

In partial Fulfillment of the Requirements

For the Degree of

Master in Statistics

June 2005

Hsinchu Taiwan Republic of China

中華民國 九十四 年 六 月

中文摘要

一般的統計分析，偏重在兩變數間的相關性，而無法表示其因果關係。例如，身高跟體重有相當顯著的正相關，是身高影響體重還是體重影響身高呢？此例中，改變一個人的體重並不會使其長高或變矮，故體重不是影響身高的原因。但在現實生活、產業製程、甚至是生物基因科技中，有許多變數間的因果關係並不是那麼顯淺易見，所關心的變數之間到底何為親輩？何為子輩？一直是科學家所想要探索的議題。而基因網路正是目前分析變數間因果關係最強而有力的工具。

簡單來說，基因網路主要由兩部分所構成：一、代表變數的節點(node)：二、連接節點之間的線段(edge)，以箭頭表示因果方向(Lauritzen, 1982；Wermuth and Lauritzen, 1983；Kiiveri et al., 1984)。

結合專業領域知識與資料，有助於基因網路的建構。基因網路最重要的特色之一，便是具有學習功能。可藉由資料的不斷更新，來學習機率模型的設定，使推估達到穩定，有效的來作機率推論的工作(Pearl, 1986)。推論是基因網路最主要的用途，近年來其應用已相當廣泛，如醫學診斷、資訊檢索、預測市場未來走向、天氣預報和人工智慧等等。而目前最熱們的基因工程，探討某基因的存在是否會活化或抑制另一基因，基因間的因果關係更是科學家所想要迫切了解的(Pe'er et al., 2001 ; Spirtes et al., 2000)。

本論文將應用兩種統計方法(PC-Algorithm & 貝氏計分法)，探討變數間之因果關係並建構基因網路之模型。

Abstract

If we have several random variables, the statistician usually focus on the joint distribution between variables. But when two variables are highly correlated does not mean that one causes the other. In statistical term, we say that correlation does not imply causation.

Over the last decade, researchers have been developed many methods for inferring causality. In this paper, we describe a Gene Network as an efficient tool for causality discovery by fusion of data and prior knowledge.

A Gene Network is a graphical model that encodes probabilistic relationships among variables of interest. Recently, the Gene Network has become a popular representation for encoding uncertain expert knowledge in expert system. They have been used to aid in the diagnosis of medical patients and malfunctioning systems, to filter documents, to facilitate planning in uncertain environments. But the construction of a Gene Network can be a very difficult task to perform. For this reason, we discuss two methods: (1)PC-Algorithm, (2)Bayesian Score, for constructing Gene Network.

誌謝

本論文得以順利完成，首先要感謝恩師 洪慧念教授的教導與鼓勵。

承蒙交通大學 陳志榮教授、中央研究院統計科學研究所 黃信誠教授與中央研究院統計科學研究所 謝淑蓉教授對本論文不吝指正，提供許多精闢見解與改進諸項缺失，僅此表達誠摯的謝枕。

論文研究期間，感謝統計學研究所諸位教授之教導與各位教職員的協助，以及研究室伙伴碰仔、景嵐、辰昀、玉均、雅靜、如美等，平日在學業與生活上的照顧。要感謝的人太多，請原諒無法一一列出，我把這份感謝銘記於心。

感謝父親 林安平先生、母親 陳香女士、弟弟以及所有親人，他們無止息的關懷與代禱是支持我奮發向上、克服困難的動力。最後特別感謝女友宜敏的體貼與關懷，陪我走過這一段艱辛的過程，讓我得以盡心完成學位。

僅將本論文獻給各位。

目錄

圖目錄

表目錄

第一章　導論

基因網路是一種以圖形模式來呈現變數間因果關係的方法。在最近十幾年來，基因網路已成為一種流行，用在專家系統中來解碼不確定性的專家知識。近來，已發展了很多藉由資料來學習基因網路的方法。這些技巧至今還在持續發展當中，在資料分析問題上，它們已顯現出卓越的效能了(Pearl, 1995；Shafer and Pearl, 1990；Shachter, 1990；Oliver and Smith, 1990；Neapolitan, 1990)。

1-1 何謂因果(Causal & Effect)

長久以來，因果問題一直受到爭議，對於因果似乎一直都沒有一個明確的定義。Constantin F. Aliferis & Ioannis Tsamardinos (2003)在" Discovery of Causal Structure Using Causal Probabilistic Networks Induction"一文以機率觀點，對因果做出下列較理想的定義：

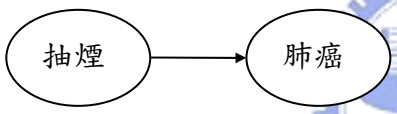
- Assume the existence of a mechanism M capable of setting values for a variable A. We say that A can be manipulated by M to take the desired values.
- Variable A causes variable B, if: in a hypothetical randomized controlled experiment in which A is randomly manipulated via M (i.e., all possible values ai of A are randomly assigned to A via M) we would observe in the sample limit that $P(B= b \mid A= ai) \neq P(B= b \mid A=aj)$ for some $i \neq j$.

1-2 因果(causality)和統計相關(association)之關係

統計分析偏重在兩變數間的相關性，當兩變數為顯著相關時，我們想瞭解其間因果關係為何，是直接因果、間接因果、還是毫無因果關係。舉例說明，血壓與收入，由資料可發現血壓與收入有正相關，即高血壓有高收入。但事實上，改變某人血壓的高低並不會影響收入的多寡，反之亦然。因此血壓與收入間並沒有因果關係，它們均受到第三個變數〝年齡〞的影響。年齡大的人收入較多但血壓也較高，年齡小的收入較低但有較佳的健康狀態，所以血壓較低。再以圖形和一簡單例子來表示因果和統計相關，有下列三種情形：

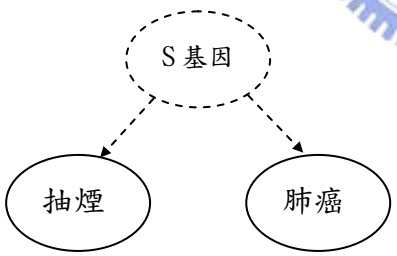
(事實：抽煙和肺癌有顯著正相關，S 基因為隱藏變數)

(i)



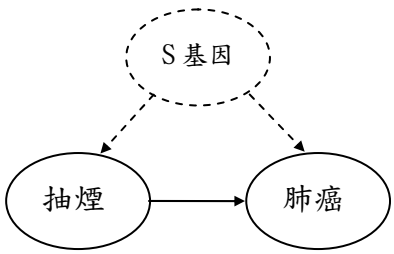
抽煙造成肺癌

(ii)



S 基因使人容易抽煙亦容易罹患肺癌

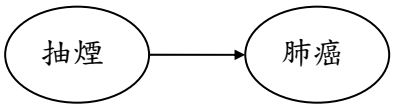
(iii)



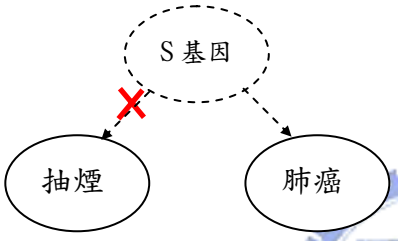
S 基因使人容易抽煙亦容易罹患肺癌，且抽煙亦會造成肺癌

在上述三種情況下，在只有抽煙和肺癌的資料時，統計上，均可呈現出顯著相關，顯而易見地，僅有在(i)、(iii)情況下，抽煙和肺癌才有真正的因果關係。若在本實驗中，可控制抽煙情形，即去除 S 基因影響抽煙的成分，此時，在(ii)狀態下，抽煙與肺癌將不再相關。如下圖：

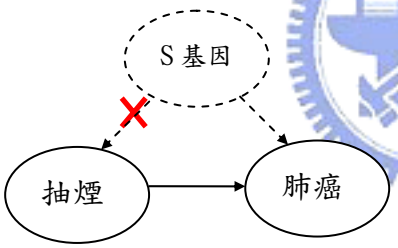
(i)



(ii)



(iii)



因此，有統計相關不一定為因果關係，然而存在因果關係之變數必定統計相關。

若來源資料為不可操控(uncontrollable variable)或有隱藏變數(hidden variable)的資料時，如何從這些資料去察覺變數間的因果關係，是目前蠻熱門的議題。而基因網路(Gene Network)便是要將變數間的相關找出其正確的因果關係。

1-3 貝氏法則

在統計上，貝氏方法和古典機率方法有著很大的差異。在古典方法中，參數 θ 被視為未知數但為一固定的值。$X_1,\ldots,X_n$ 是來自母體的隨機樣本，藉由在樣本所觀察到的值，可以獲得關於母體參數 θ 的資訊。而在貝氏方法中，θ 被視為一個量，其變異可以被一機率分配 $P(\theta)$（即為先驗分配(prior distribution)）所描述。這是一主觀的分配，建構在實驗者本身的信念，並且是在資料被看見之前所形成，因此稱為先驗分配。一來自母體的樣本，先驗分配可以藉由樣本 D 所提供的資訊而獲得更新。此更新的先驗便稱為後驗分配 $P(\theta \mid D)$ (posterior distribution)。而更新的動作可靠貝氏規則(Baye's Rule)完成。過程如下：

$$\underset{posterior}{P(\theta \mid D,\kappa)} = \frac{\overset{prior}{P(\theta \mid \kappa)}\overset{likelihood}{P(D \mid \theta,\kappa)}}{P(D \mid \kappa)} \qquad (1.1)$$

其中 θ 為母體參數、D 為給定的資料、κ 則為背景知識。

本研究所欲探討的基因網路，便是建構在貝氏機率下所完成的。

第二章　文獻探討

2-1 基因網路之結構

簡單的說，基因網路是由下列三部分所構成：

一、節點(nodes)：

即所欲探討的變數(variables)。

二、連接節點的有向線段(edges)：

顯示變數之間的因果關係。

三、參數(parameters)：

即在圖形 G 中所對應變數的條件機率，可表示成 $P(X_i \mid Pa_i^G)$，其中 Pa_i 表 X_i 的親輩(parents)，而 X_i 的變化在圖形 G 中只受其親輩 Pa_i 的影響。

由此三部分所構成的一個有向非環狀圖形(directed acyclic graph，簡稱 DAG)即為基因網路。並利用親屬關係之術語（如：親輩(parents)、子輩(children)、配偶(spouses)、子孫(descendants)、祖先(ancestors)），來表示圖形中變數之關係。

由貝氏觀點，聯合機率分配可分解成

$$P(X_1,...,X_n) = P(X_n \mid X_1,...,X_{n-1}) \cdot P(X_{n-1} \mid X_1,...,X_{n-2}) \cdot ... \cdot P(X_1) \text{ 。} \quad (2.1)$$

因在基因網路中 X_i 只受其親輩 Pa_i 影響，利用條件獨立(conditional independence)之概念(稍後定義)，可將聯合機率分配簡化成

$$P(X_1,...,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_i) \text{ 。} \quad (2.2)$$

定義：條件獨立(conditional independence)

V={V1,V2,…,Vn}為變數集合。P(・)為V的聯合機率分配，設X、Y、Z為V中的任意子集。若$P(x \mid y, z) = P(x \mid z)$ 則稱，在給定Z下，X和Y條件獨立。記為 $X \perp Y \mid Z$。

舉例說明，一社區中之警報器(A)，在有夜賊(B)出沒和地震(E)發生時會啟動，此時將有民眾報警(C)。若當地震發生，則社區廣播器(R)亦會發出警示。因此，基因網路圖可表示為：

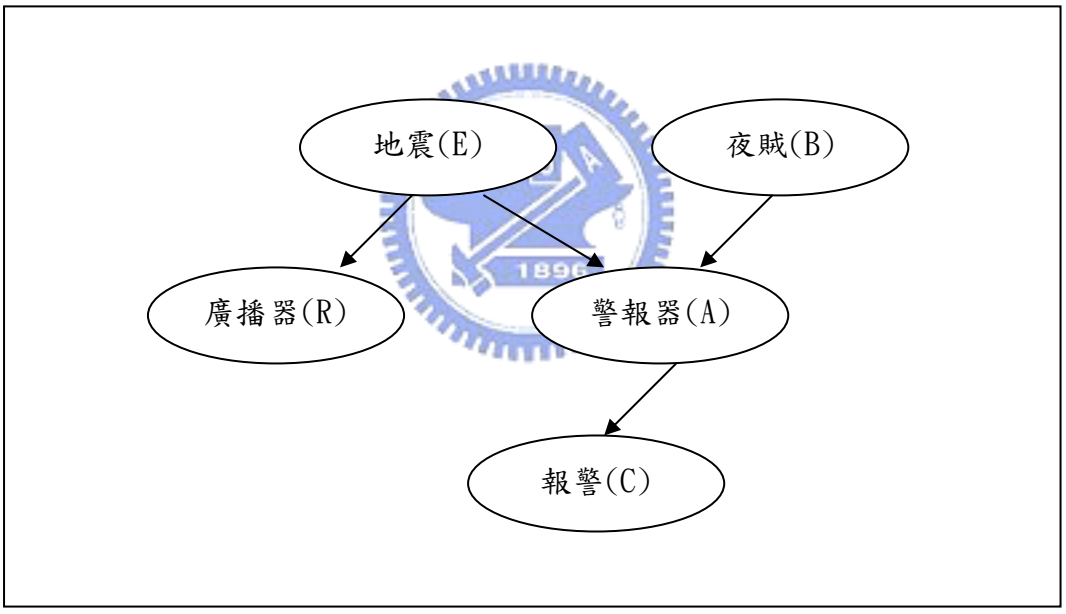


圖 2.1 基因網路圖

由(2.1)，故聯合機率分配可表示為

$$P(C, A, R, E, B) = P(C \mid A, R, E, B)P(A \mid R, E, B)P(R \mid E, B)P(E \mid B)P(B)$$

。在條件獨立下，可簡化為

$$P(C, A, R, E, B) = P(C \mid A)P(A \mid E, B)P(R \mid E)P(E)P(B)$$ 。

2-2 基因網路之性質

2-2.1 因果假設(Causal Assumption)

實際上，因果結構內變數是無法真正完全找出來的，故要探討變數 X 到變數 Y 之間的因果關係可能會有很多種模型，而每一種模型所涉及的隱藏變數亦有所不同，所以不同的模型便會造成不同的因果關係出現，所以當我們要搜尋一個適當的 DAG 時，是必須有所限制的(Spirtes et al., 2000)。

(i).因果充分性假設(Causally Sufficient Assumption)：

在可觀察的變數集合 V 中，其中任兩變數都沒有存在著隱藏性的共同原因(hidden common causes)。

(ii).馬可夫因果假設(Causal Markov Assumption)：

在給定其親輩下，每一變數都跟其非子輩(non-descendants)的變數獨立。

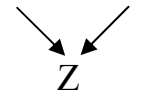
2-2.2 d-separation & V-Structure (Verma and Pearl , 1988)

基因網路中另一個重要的特色就是 d-separation 和 V-Structure。Pearl 在 ”CAUSALITY ”(2000) 一書定義如下：

Definition (d-Separation) A path p said to be d-separated (or blocked) by a set of nodes Z if and only if 1. p contains a chain i→m→j or a fork i←m→j such that the middle node m is in Z , or 2. p contains an inverted fork (or collider) i→m←j such that the middle node is not in Z and such that no descendant of m is in Z.

簡單的說，d-separation 的出現會打斷兩變數間的因果相關，使其變為條件獨立。以圖形表示：（一）X→Z→Y（二）X←Z→Y（三）X←Z←Y，其中 Z 均為 d-separation。此三種圖形，在給定 Z 之情況下，X 和 Y 均條件獨立。

Definition (V-Structure)
An ordered triple of nodes (X , Y , Z) in a graph such that
1. X→Z and Y→Z
2. X and Y are not adjacent (i.e., not connected by an edge)



由圖形可知，V-Structure 即為配偶與子代之關係(兩變數共同影響第三變數)。在給定子代的情形下，使原本獨立之兩變數，將會變為相關。

藉由 d-separation 和 V-Structure 的瞭解，探索結構中條件獨立之關係，有助於將來建構模型時對變數間因果方向之確認。

2-2.3 等價關係(equivalent)

若當兩個 DAG 呈現出相同的獨立關係時，此時稱此兩圖形等價(equivalent)(Verma and Pearl , 1990)。如 X→Z→Y 、 X→Z→Y 及 X←Z→Y 均表示在給定 Z 的情況下，X 和 Y 獨立。此時從觀察的資料看來，我們並無法確認出何者才為正確的 DAG，但若有時間先後的資訊、經驗累積或專業知識，或許可區分出部分的 DAG。因此我們想去探索某些共同的性質在這些等價的 DAG 中。在這些等價 DAG 中可以表示成一個唯一的部分有向非環狀圖形(partially directed acyclic graph，簡稱 PDAG)。在 PDAG 中有些變數的關係可以被確認出，有些則否。

2-3 建構基因網路

本研究將使用下列兩種方法來建構基因網路。

2-3.1 PC-Algorithm (Spirtes et al., 1993)

此法主要建立在變數間的獨立性檢定，藉由這些獨立關係，PC-Algorithm 將會建構出等價的基因網路圖形(PDAG)。本節將介紹兩種無母數的檢定方法，來探討變數間的獨立關係。

獨立性檢定：

(i)Kendall's Tau (Siegel, 1956)

假設有 N 組觀察值(x_i, y_i)，i=1,…,N，因此可產生$C_2^N = \frac{1}{2}N(N-1)$不同之配對數。設$(x_i, y_i)$和$(x_j, y_j)$為一配對觀察值，若$(x_i - x_j)(y_i - y_j) > 0$，稱此配對一致性(concordant)；$(x_i - x_j)(y_i - y_j) < 0$，稱此配對不一致(discordant)。令 C 為一致性之配對數；D 為不一致之配對數，則 Kendall's Tau 被定義為：

$$\tau = \frac{C - D}{C_2^N} \text{。} \tag{2.3}$$

若 $x_i = x_j$ 或 $y_i = y_j$ 此種配對個數過多，則(2.3)將被調整為：

$$\tau = \frac{C - D}{\sqrt{(C_2^N - N_x)(C_2^N - N_y)}}, \tag{2.4}$$

其中 N_x 為 $x_i = x_j$ 之配對個數，N_y 為 $y_i = y_j$ 之配對個數。

當兩變數 X、Y 獨立時，由定義可直覺地發現，τ 值會被期望為 0。當 $N \geq 10$ 時，τ 將會快速地逼近常態分配，且變異數為 2(2N+5)/9N(N-1)。故檢定統計量可設為：

$$z = \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} \text{。} \tag{2.5}$$

因此檢定兩變數獨立時，只需觀察 τ 值是否有落在 $\pm 1.96[2(2N+5)/9N(N-1)] \fallingdotseq \frac{4}{3}n^{\frac{-1}{2}}$ 界線之外(假定型一誤差 α 為 5%)，有則棄卻獨立假設，反之則否。

(ii)Spearman's Rank Correlation (Siegel and Castellan, 1988 ； Siegel, 1956)

假設有 N 組觀察值(x_i , y_i)，i=1,…,N。分別對 X 和 Y 排序，即 rank(min_x_i)=1、rank(max_x_i)=N。令 $R_i = rank(x_i)$ 且 $S_i = rank(y_i)$，則 Spearman's rho 定義如下：

$$r_s = \frac{\sum(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum(R_i - \overline{R})^2}\sqrt{\sum(S_i - \overline{S})^2}} = \frac{\{\sum R_i S_i - N(N+1)^2/2\}}{\{N(N^2-1)/12\}} = \frac{1 - 6\sum(R_i - S_i)^2}{\{N(N^2-1)\}} \text{。} \tag{2.6}$$

當 N≥10 時，定義檢定統計量為：

$$t = r_s\sqrt{\frac{N-2}{1-r_s}} \text{。} \tag{2.7}$$

在虛無假設 H_0：$r_s = 0$ 下，此統計量 t 將會近似於 t 分配且自由度為 N-2。

PC-Algorithm 主要步驟如下：

Step1：連接變數集合 V 中任兩相異變數。

Step2：在變數集合 V 中，對每一對變數 a、b，檢驗其(條件)獨立關係。搜尋是否存在一子集 S_{ab} 使得 a、b 獨立(即 $a \perp b \mid S_{ab}$)。若一旦存在 S_{ab}，則移去 a、b 間之線段。(i.e.,判斷是否存在 d-separation)

Step3：利用所檢驗出之(條件)獨立關係，建構一無向性(undirected)之基因網路。

Step4：對任兩不相鄰變數 a、b 且存在共同相鄰變數 c(即 $a-c-b$)，判別 c 是否屬於 S_{ab}。若 $c \notin S_{ab}$，則 a→c←b。(i.e.,判斷是否存在 V-Structure)

Step5：在限制條件：(i)不可創造出新的 V-Structure；(ii)不可產生有向循環(directed cycle)，和利用定向規則(Verma and Pearl , 1992)(稍後說明)，建構出一最多方向性的 PDGA。

定向規則：

R_1：Orient b-c into b→c whenever there is an arrow a→b such that a and c are not adjacent.

R_2：Orient a-b into a→b whenever there is a chain a→c→b.

R_3：Orient a-b into a→b whenever there are two chains a-c→b and a-d→b such that c and d are nonadjacent.

R_4：Orient a-b into a→b whenever there are two chains a-c→d and c→d→b such that c and b are nonadjacent.

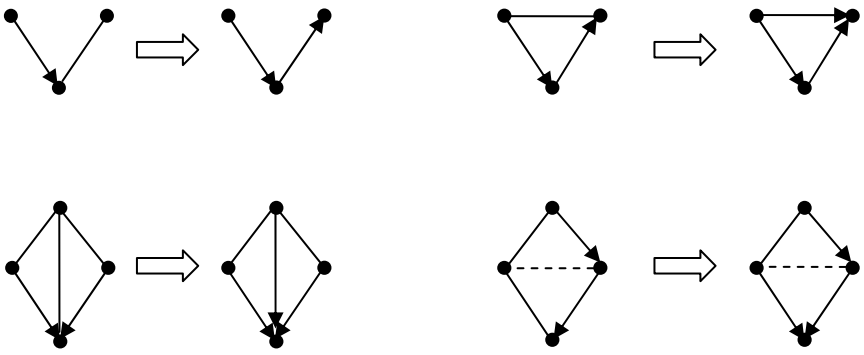


圖 2.2 四個定向規則

將上述步驟簡化成演算法模式：

1.Start with a complete undirected graph GP
2. $i = 0$
3.Repear
  4.For each $X \in V$
    5.For each $Y \in ADJ_x$,where ADJ_x is adjacent node of X
      6.Determine if there is $S \subseteq ADJ_x - \{Y\}$ with $|S| = i$ and $X \perp Y \mid S$
      7.If this set exists
        8.Make $S_{XY} = S$
        9.Remove X-Y
  10. $i = i + 1$
11.Until $|ADJ_X| \le i, \forall X$

(演算法摘錄自：Seraf'in Moral "Structure Learning : The PC-Algorithm" , 2003)

以圖示說明 PC-Algorithm 之建構過程：

變數集合 V＝{A,B,C,D,E}

Step1：建構出完整之結構

正確模型

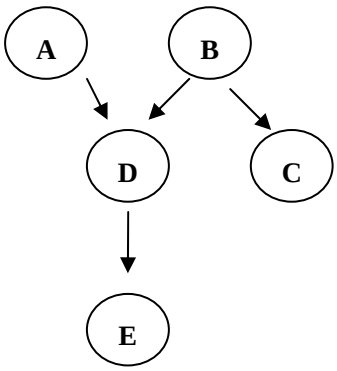


圖 2.3 正確之基因網路結構

連接 V 中任兩相異變數

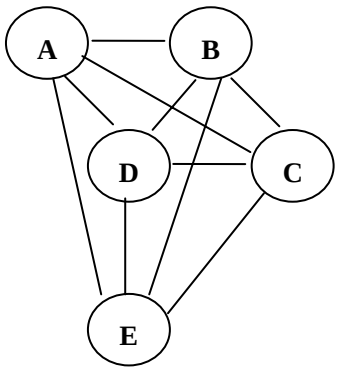


圖 2.4 完整無向性之結構

Step2：判斷是否存在 d-separation

➔ $A \perp B, A \perp C,$
$A \perp E \mid D, B \perp E \mid D, C \perp D \mid B, C \perp E \mid B$

➔除去 A-B,A-C,A-E,B-E,C-D,C-E 之線段

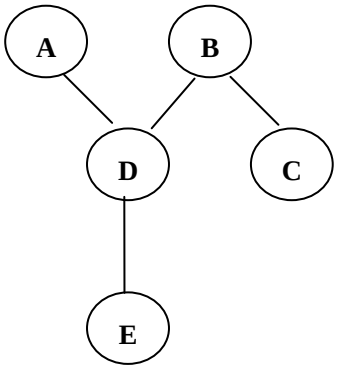


圖 2.5 無向性結構

Step3：判斷是否存在 V-Structure

➔ $D \notin S_{AB} \Rightarrow A \rightarrow D \leftarrow B$

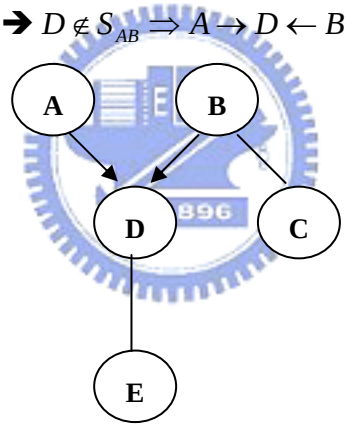


圖 2.6 部分有向非環狀圖形

Step4：利用定向規則建構出最終 PDGA

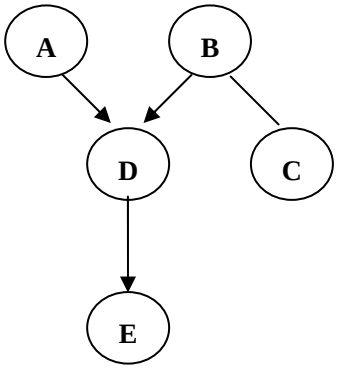


圖 2.7 最大部分有向非環狀圖形

PC-Algorithm 可建構出一 PDGA。如 2-2.3 節所言，對於圖形中某些未確定的方向，可藉由時間先後或背景知識的條件來確認之。若所處理之變數為可控制變數，則可利用干預法(intervention)有效地確認出因果關係之方向。以圖 2.7 說明，$D \leftarrow B - C$ 中的關係，可能為 (i) $D \leftarrow B \rightarrow C$ 或(ii) $D \leftarrow B \leftarrow C$。此時設定 C 為所操控變數，則 C 將不再受其親輩影響，但亦會繼續影響子輩。故若實驗中，改變 C，B 不為所動，則為(i)之情形；改變 C，B 隨之變動，則為(ii)之情形。

2-3.2 貝氏計分法(Score) (Heckerman et al. , 1996)

此為一 Bayesian 方法，在給定資料 D 下，對於其對應的每一個圖形 G，求出其後驗機率(posterior probability) $P(G \mid D)$。定義貝氏計分：

$$S(G:D) = \log P(G \mid D) = \log \frac{P(D \mid G)P(G)}{P(D)} = \log P(D \mid G) + \log P(G) + C \text{。} \tag{2.7}$$

其中 $P(D \mid G)$ 為概似機率(likelihood)，$P(G)$ 為先驗機率(prior)，C 為一常數，與 G 的選取無關。所以當有越高的後驗分配，也就是在已知的資料下支持此圖形 G 的機率越高，則有越高的貝氏計分 $S(G:D)$。且

$$P(D \mid G) = \int_{\theta} P(D \mid G, \theta) P(\theta \mid G) d\theta \text{。} \tag{2.8}$$

其中 θ 為對應圖形 G 中所涉及之參數。由(2.7)、(2.8)可發現，G 的先驗機率函數 $P(G)$ 與 $P(\theta \mid G)$ 的選取，會直接影響圖形 G 的貝氏計分 $S(G:D)$。

此法與 PC-Algorithm 最大的差異在於需要了解變數的分配

(distribution)，以便於計算概似函數$P(D|G)$。而貝氏計分法另一面臨的挑戰，便是圖形的搜尋。圖形的搜尋，其複雜度為一 NP-hard 的問題(Cooper, 1990)。以三變數為例，不考慮因果方向，圖形結構有以下八種可能：

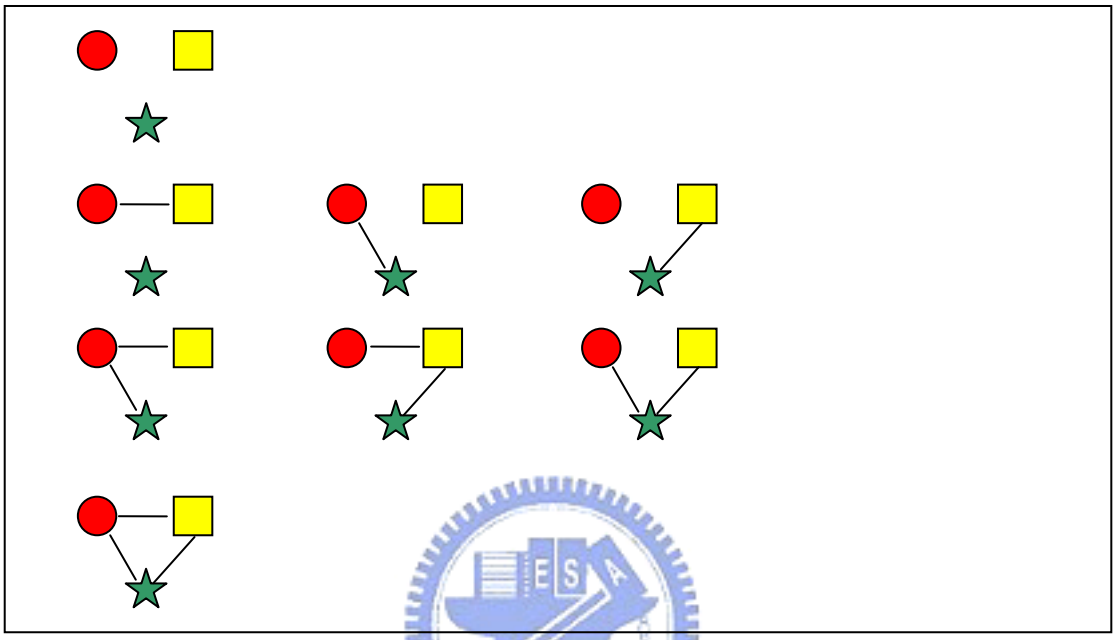


圖 2.8 三變數之八種結構圖

加入因果方向，此八種結構便會產生 25 個相異之 DAGS。若有N個變數，則有$2^{N(N-1)/2}$種不同的結構圖形(尚未考慮因果方向)。一旦變數個數增加，圖形個數即呈指數倍成長，龐大的圖形空間將會使計算趨於複雜、效率低落。而結構中線段越多，則參數數目越多，因此可能會產生過度配適(overfit)，使不正確圖形卻有較高的貝氏計分。因此如何有效的簡化圖形空間和選擇適當的先驗機率$P(G)$，都是目前值得研究的議題。

第三章　模擬變數，建構基因網路

3-1 模擬變數，利用 PC-Algorithm 建構基因網路

3-1.1 探討變數之樣本數、誤差大小與β之關係

PC-Algorithm 已被證實，在提供無限多組資料之條件下，才能建構正確的基因網路(Madsen et al. , 2003)。在真實生活中，資料卻是有限的。一般而言，在有限資料下，欲求出變數間所有真實的(條件)獨立關係是非常困難的，尤其在當資料稀少或變數之測量誤差大的情況下，往往會檢定出多餘的(條件)獨立關係(即易產生較大的型二誤差β)，而造成錯誤的基因網路模型。因此利用統計檢定所導出的 DAGs，與真實的 DAGs 相較，可呈現出相同的(條件)相關，但無法表現所有相同之(條件)獨立關係。

本節欲利用 Kendall's tau，檢定變數間之獨立關係，並探討變數之樣本數、誤差大小與β之關係。

(設定型一誤差 α＝0.05)

一、當誤差為變數的兩倍時，樣本數與β的關係

當資料如下時：Y=X+2ε，where X and ε are iid N(0,1)

做下面之 test：H_0：X、Y 獨立　VS.　H_1：X、Y 相關

表 3.1：Y=X+2ε，樣本數與β的關係

樣本數	重複次數	Reject H_0 次數	型二誤差β
50	1000	839	0.151
80	1000	976	0.024
100	1000	997	0.003
150	1000	1000	0

二、當誤差為變數的十倍時，樣本數與β的關係

當資料如下時：Y=X+10ε，where X and ε are iid N(0,1)

做下面之 test：H_0：X、Y 獨立　VS.　H_1：X、Y 相關

表 3.2：Y=X+10ε，樣本數與β的關係

樣本數	重複次數	Reject H_0 次數	型二誤差β
50	1000	108	0.892
100	1000	152	0.848
500	1000	579	0.421
1000	1000	864	0.136
1500	1000	955	0.045

三、當樣本數＝100 時，誤差大小與β的關係

當資料如下時：Y=X+Aε，where X and ε are iid N(0,1)

做下面之 test：H_0：X、Y 獨立　VS.　H_1：X、Y 相關

表 3.3：Y=X+Aε，樣本數與β的關係

A	重複次數	Reject H_0 次數	型二誤差β
10	1000	150	0.850
5	1000	452	0.548
2	1000	992	0.008
1	1000	1000	0

四、當樣本數＝100 時，模型與β的關係

當資料如下時：X and ε are iid N(0,1)　樣本數 N＝100

做下面之 test：H_0：X、Y 獨立　VS.　H_1：X、Y 相關

表 3.4：樣本數＝100 時，模型與β的關係

模型	重複次數	Reject H_0 次數	型二誤差β
Y=X+ε	1000	1000	0
Y=X*ε	1000	136	0.864
Y=X*(0.2+ε)	1000	556	0.444
Y=X*(0.5+ε)	1000	991	0.009
Y=X*(1+ε)	1000	1000	0

由上述模擬結果可發現，當樣本數少(約小於 50)或誤差大(約大於所研究之變數 5 倍)的情況下，確實有較大的型二誤差β，故易造成模型誤判的現象。為了避免錯誤模型的產生，應可考慮型一誤差與型二誤差之平衡。

由於較小的樣本數易導致型二誤差β的上升，故對於小樣本，我們或許可考慮採取較低之門檻，即容許較高之型一誤差α，以降低型

二誤差β之機率。其模擬結果如下：(表格內灰色區域為β值)

表 3.5：在模型 Y=X+ε下，樣本數與α所導致之β

模型 Y=X+ε		型一誤差 α			
		0.05	0.10	0.15	0.20
樣本數	100	0	0	0	0
	50	0	0	0	0
	20	0.089	0.039	0.028	0.015
	10	0.435	0.280	0.228	0.167

表 3.6：在模型 Y=X+2ε下，樣本數與α所導致之β

模型 Y=X+2ε		型一誤差 α			
		0.05	0.10	0.15	0.20
樣本數	100	0.007	0.004	0.003	0.001
	50	0.116	0.057	0.038	0.028
	20	0.536	0.395	0.317	0.259
	10	0.796	0.648	0.560	0.476

表 3.7：在模型 Y=X+5ε下，樣本數與α所導致之β

模型 Y=X+5ε		型一誤差 α			
		0.05	0.10	0.15	0.20
樣本數	100	0.552	0.427	0.342	0.262
	50	0.747	0.642	0.561	0.468
	20	0.871	0.792	0.749	0..656
	10	0.929	0.844	0.783	0.717

由上列表可發現，放寬型一誤差α確實可有效降低β之機率，但當誤差過大時，仍有偏高之β值，而無法建構正確基因網路。

3-1.2 模擬一結構方程式，建構基因網路

基因網路中，變數x_i只受其親輩pa_i之影響，因此模型可以方程式表示為：

$$x_i = f(pa_i, \varepsilon_i), \quad i = 1,...,n. \tag{3.1}$$

其中ε_i為造成x_i之誤差的原因，亦稱干擾變數，可能來自於所忽略的隱藏變數或測量之偏差。(3.1)式稱之為結構方程式(Structural Equations)(Peral, 2000)。本節欲模擬一結構方程式，並利用PC-Algorithm建構基因網路。

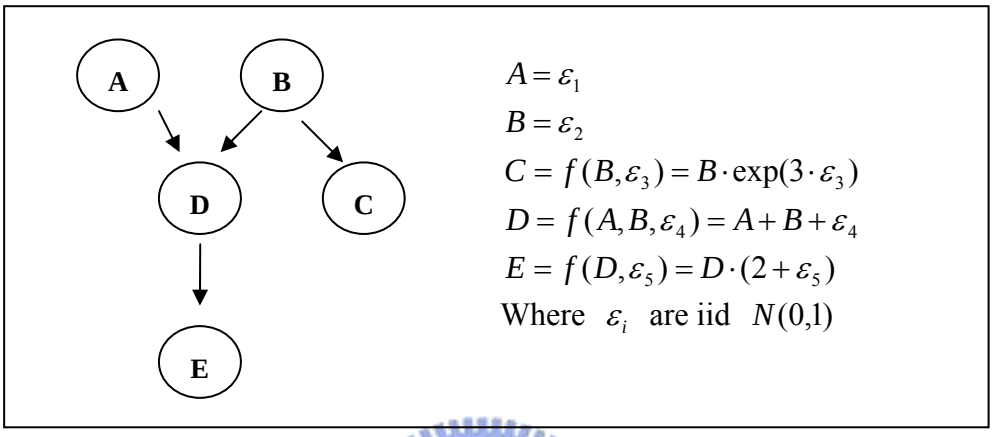


圖 3.1 基因網路與其結構方程式

模擬結果如下：(樣本數 N＝100，型一誤差 α＝0.05)

（一）獨立性檢定

$$H_0 : A \perp B \quad \text{VS} \quad H_1 : A \not\perp B$$

表 3.8：獨立性檢定

變數	τ 值	統計量 z 值	P-Value	結論
A、B	0.024	0.354	0.723	獨立
A、C	0.04	0.59	0.555	獨立
A、D	0.375	5.528	3E-08	相關
A、E	0.333	4.909	9E-07	相關
B、C	0.624	9.199	0	相關
B、D	0.396	5.838	5E-09	相關
B、E	0.377	5.558	3E-08	相關
C、D	0.324	4.776	2E-06	相關
C、E	0.308	4.54	6E-06	相關
D、E	0.781	11.51	0	相關

由表 3.8 可知，在此模型下，兩兩變數之間的關係有很大的機率可被完全正確地確認出來。

（二）條件獨立性檢定

對於條件獨立檢定問題，本論文欲利用兩種不同方法考慮之：(i)修正之 p-value；(ii)修正之型一誤差 α^*。

(part1 :修正之 p-value)

欲求給定 C 變數，A、B 間之獨立關係，首先將給定變數 C 分組，利用 $Int[\frac{Rank(C_i)}{N/k}]$ 進行分組，其中 N 為 C 之樣本數、k 為分組數。分別對 k 組內之對應變數 A、B 進行 Kendall's tau 獨立性檢定，最後檢視各組內之 p-value。因分組之緣故，每組內之樣本數為 N/k，為避免縮小的樣本而導致 β 的增加，因此對組內所計算的 p-value 稍做修正。將分組內之 p-value 取自然對數 ln，由於 p-value 介於 0～1，且當 p-value 越小時，則所對應之自然對數會越快速遞減，故此轉換動作，對較小的 p-value 有加權之效果。將 k 組 p-value 所轉換的自然對數加總並取平均值，以最後的平均值作為指數的次方，其值即為修正後之 p-value。

模擬結果如下：(樣本數 N＝100，分組數 k＝5)

H_0：A⊥B｜C　VS　H_1：A⊥̸B｜C (註：若 p-value≤0.001 以 0.001 記之)

表 3.9：條件獨立性檢定(part1)

變數	給定變數	p-value					修正之 p-value	結論	是否誤判
		第一組	第二組	第三組	第四組	第五組			
A、B	C	0.43	0.43	0.85	0.81	0.53	0.583	獨立	正確
	D	0.64	0.001	0.07	0.46	1	0.115	獨立	誤判
	E	0.67	0.001	0.49	0.11	0.5	0.112	獨立	誤判
A、C	B	0.68	0.24	0.79	0.67	0.31	0.484	獨立	正確
	D	0.57	0.001	0.27	0.05	0.84	0.091	相關	正確
	E	0.41	0.001	0.83	0.8	0.67	0.178	獨立	誤判
A、D	B	0.08	0.001	0.01	0.06	0.001	0.008	相關	正確
	C	0.001	0.02	0.01	0.001	0.01	0.004	相關	正確
	E	0.001	0.59	0.06	0.04	0.001	0.016	相關	正確
A、E	B	0.18	0.001	0.07	0.08	0.001	0.015	相關	正確
	C	0.01	0.02	0.03	0.001	0.04	0.011	相關	正確
	D	0.07	0.89	0.87	0.39	0.03	0.229	獨立	正確
B、C	A	0.001	0.001	0.001	0.001	0.001	0.001	相關	正確
	D	0.14	0.001	0.001	0.01	0.06	0.009	相關	正確
	E	0.03	0.001	0.001	0.07	0.08	0.011	相關	正確
B、D	A	0.001	0.001	0.001	0.001	0.001	0.001	相關	正確
	C	0.84	0.03	0.01	0.14	0.32	0.102	相關	正確
	E	0.05	0.57	0.11	0.86	0.02	0.140	獨立	誤判
B、E	A	0.05	0.001	0.001	0.01	0.001	0.003	相關	正確
	C	0.9	0.7	0.001	0.37	0.53	0.165	獨立	誤判
	D	0.76	0.68	0.06	0.7	0.16	0.322	獨立	正確
C、D	A	0.01	0.001	0.06	0.14	0.001	0.009	相關	正確
	B	0.74	0.2	0.31	0.72	0.62	0.459	獨立	正確
	E	0.02	0.88	0.42	0.56	0.95	0.330	獨立	誤判
C、E	A	0.08	0.001	0.11	0.06	0.001	0.013	相關	正確
	B	0.98	0.28	0.28	0.83	0.61	0.522	獨立	正確
	D	0.13	1	0.22	0.89	0.26	0.366	獨立	正確
D、E	A	0.001	0.001	0.001	0.001	0.001	0.001	相關	正確
	B	0.001	0.001	0.001	0.001	0.001	0.001	相關	正確
	C	0.001	0.001	0.001	0.001	0.001	0.001	相關	正確

由表 3.9 可知，所有檢定為條件相關之關係均正確，但會誤判出多餘的條件獨立之關係(此例中有六個誤判)，而所誤判之條件獨立關係大部分有較小的 p-value，約介於 0.1～0.2。

整合上述(條件)獨立檢定結果：

$A \perp B, A \perp C$

$A \perp B \mid C, A \perp B \mid D, A \perp B \mid E, A \perp C \mid B, A \perp C \mid E, A \perp E \mid D, B \perp D \mid E,$
$B \perp E \mid C, B \perp E \mid D, C \perp D \mid B, C \perp D \mid E, C \perp E \mid B, C \perp E \mid D$

由 PC-Algorithm 可知，當存在一子集 S，使得 $X \perp Y \mid S$，則除去 X、Y 之線段。因此除去 $A-B, A-C, A-E, B-D, B-E, C-D, C-E$ 線段。且 $A \perp E \mid D$，因此結構中並無 V-Structure，故無法確認出任何方向(見圖 3.2)。

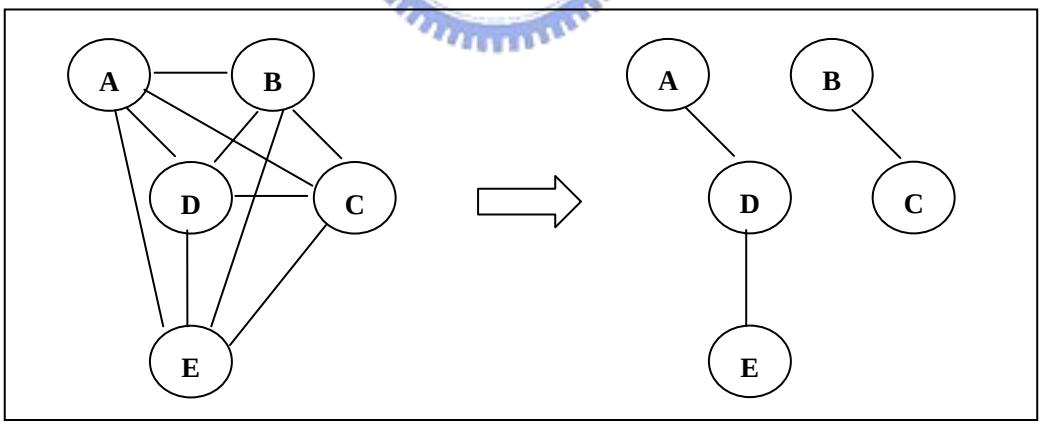


圖 3.2 PDAG

由圖 3.2 發現，錯誤的檢定結果，不僅使 PC-Algorithm 的效果不彰，導致出不正確之 PDAG，且其 PDAG 與檢定出之(條件)獨立關係有多處矛盾之處，如圖形顯示出 B 與 D、E 獨立，C 與 D、E 獨立，$A \perp C \mid D$ 等等，均與檢定的結果不符。

錯誤的模型，來自於錯誤的(條件)獨立檢定結果，其主要原因則是因條件獨立性檢定偏高的型二誤差 β 所造成，易使條件相關為真的情形下，誤判為條件獨立，因此我們將焦點放在檢定出的條件獨立關係上，本論文欲以某些準則，設法刪除部分條件獨立之關係，重新建構一較合理、正確之基因網路模型。

由表 3.9 所檢定出的條件獨立關係如下：

$$A \perp B \mid C, A \perp B \mid D, A \perp B \mid E, A \perp C \mid B, A \perp C \mid E, A \perp E \mid D, B \perp D \mid E,$$
$$B \perp E \mid C, B \perp E \mid D, C \perp D \mid B, C \perp D \mid E, C \perp E \mid B, C \perp E \mid D$$

在這些條件獨立關係中，可發現三組互相矛盾之關係（附錄一）：

(1) $B \perp D \mid E$ (p-value＝0.140)、$B \perp E \mid D$ (p-value＝0.322)

(2) $B \perp E \mid C$ (p-value＝0.165)、$C \perp E \mid B$ (p-value＝0.522)

(3) $C \perp D \mid E$ (p-value＝0.330)、$C \perp E \mid D$ (p-value＝0.366)

在這三組矛盾之關係中，棄卻 p-value 較小之關係，因此，

(1) $B \perp E \mid D$，則 $B \not\perp D \mid E$

(2) $C \perp E \mid B$，則 $B \not\perp E \mid C$

(3) $C \perp E \mid D$，則 $C \not\perp D \mid E$

捨棄矛盾之關係，新的(條件)獨立關係如下：

$$A \perp B, A \perp C$$

$$A \perp B \mid C, A \perp B \mid D, A \perp B \mid E, A \perp C \mid B, A \perp C \mid E, A \perp E \mid D, B \perp E \mid D,$$
$$C \perp D \mid B, C \perp E \mid B, C \perp E \mid D$$

再次利用 PC-Algorithm，除去 $A-B, A-C, A-E, B-E, C-D, C-E$ 線段，且結構中無任何 V-Structure，可得最終之 PDAG 如下：

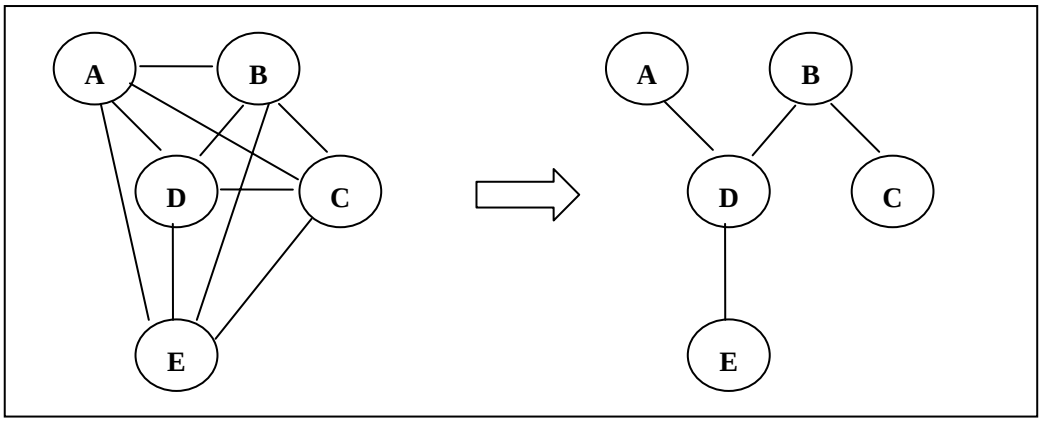


圖 3.3 PDAG

此 PDAG 和所模擬之基因網路已有相同之結構，但因仍有錯誤的條件獨立關係，而無法找到一因果方向可符合上述所有之(條件)獨立關係。

(part2 :修正之型一誤差 α^*)

對於條件獨立之檢定，我們亦可考慮另一方法。在給定變數 C 的分組中(分 k 組)，設定一修正型一誤差 α^*，若所有 k 組均無顯著證據說明 A、B 相關(即 p-value$>\alpha^*$)，則稱給定變數 C，A、B 獨立。反之，假若 k 組中，存在任一組樣本數不太小(≥20)且 p-value$\le\alpha^*$，則稱給定 C 變數，A、B 不獨立。但 α^* 應如何設定呢？令 P^* 為在 $H_0 : A \perp B \mid C$ 為真的情形下且不棄卻 H_0 之機率。則由上述方法可知，

$$P^* = (1-\alpha^*)^k \text{ 。} \tag{3.2}$$

在本節所模擬例子中，k=5 並設 $P^* = 0.90$，故 $\alpha^* = 1 - P^{1/k} = 1 - (0.90)^{1/5} = 0.02$。因此，由表 3.9 之模擬結果再判斷一次條件獨立關係，結果如下：

表 3.10：條件獨立性檢定(part2)

(註：若 p-value$\leq \alpha^* = 0.02$，以灰色區塊表示)

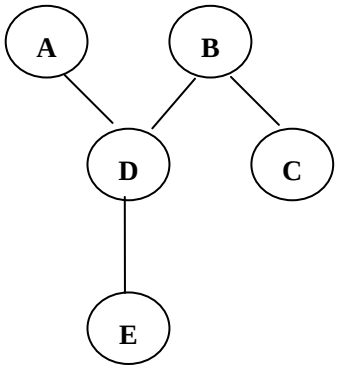
變數	給定變數	p-value					p-value $\leq \alpha^*$? ($\alpha^* = 0.02$)	結論	是否誤判
		第一組	第二組	第三組	第四組	第五組			
A、B	C	0.43	0.43	0.85	0.81	0.53	N	獨立	正確
	D	0.64	0.001	0.07	0.46	1	Y	相關	正確
	E	0.67	0.001	0.49	0.11	0.5	Y	相關	正確
A、C	B	0.68	0.24	0.79	0.67	0.31	N	獨立	正確
	D	0.57	0.001	0.27	0.05	0.84	Y	相關	正確
	E	0.41	0.001	0.83	0.8	0.67	Y	相關	正確
A、D	B	0.08	0.001	0.01	0.06	0.001	Y	相關	正確
	C	0.001	0.02	0.01	0.001	0.01	Y	相關	正確
	E	0.001	0.59	0.06	0.04	0.001	Y	相關	正確
A、E	B	0.18	0.001	0.07	0.08	0.001	Y	相關	正確
	C	0.01	0.02	0.03	0.001	0.04	Y	相關	正確
	D	0.07	0.89	0.87	0.39	0.03	N	獨立	正確
B、C	A	0.001	0.001	0.001	0.001	0.001	Y	相關	正確
	D	0.14	0.001	0.001	0.01	0.06	Y	相關	正確
	E	0.03	0.001	0.001	0.07	0.08	Y	相關	正確
B、D	A	0.001	0.001	0.001	0.001	0.001	Y	相關	正確
	C	0.84	0.03	0.01	0.14	0.32	Y	相關	正確
	E	0.05	0.57	0.11	0.86	0.02	Y	相關	正確
B、E	A	0.05	0.001	0.001	0.01	0.001	Y	相關	正確
	C	0.9	0.7	0.001	0.37	0.53	Y	相關	正確
	D	0.76	0.68	0.06	0.7	0.16	N	獨立	正確
C、D	A	0.01	0.001	0.06	0.14	0.001	Y	相關	正確
	B	0.74	0.2	0.31	0.72	0.62	N	獨立	正確
	E	0.02	0.88	0.42	0.56	0.95	Y	相關	正確
C、E	A	0.08	0.001	0.11	0.06	0.001	Y	相關	正確
	B	0.98	0.28	0.28	0.83	0.61	N	獨立	正確
	D	0.13	1	0.22	0.89	0.26	N	獨立	正確
D、E	A	0.001	0.001	0.001	0.001	0.001	Y	相關	正確
	B	0.001	0.001	0.001	0.001	0.001	Y	相關	正確
	C	0.001	0.001	0.001	0.001	0.001	Y	相關	正確

由表 3.10 可知，此法可檢定出所有正確之條件獨立關係。綜合表 3.8&表 3.10 之結果：

$A \perp B, A \perp C$

$A \perp B \mid C, A \perp C \mid B, A \perp E \mid D, B \perp E \mid D, C \perp D \mid B, C \perp E \mid B, C \perp E \mid D$

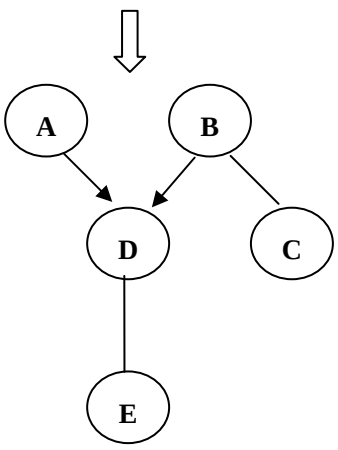
除去 $A-B, A-C, A-E, B-E, C-D, C-E$ 線段，可得圖形結構如下：



且 $A-D-B, D \notin S_{AB} \Rightarrow A \rightarrow D \leftarrow B$

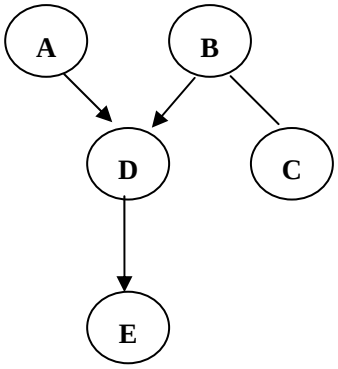
$A-D-E, D \in S_{AE} \Rightarrow A-D-E$

$C-B-D, B \in S_{CD} \Rightarrow C-B-D$



利用定向規則：

$A \rightarrow D-E \Rightarrow A \rightarrow D \rightarrow E$



最終之 PDAG

在此例中，利用修正之型一誤差α^*比修正之 p-value 所檢定出之條件獨立關係較為正確，且建構出正確之 PDAG。

由本節模擬過程可發現，利用 PC-Algorithm 建構基因網路可能遭遇到的難處：

(1) 若資料數不多，可確定出的相關性不強。

(2) 條件獨立性的檢定不易，易將條件相關誤判為條件獨立。

(3) 圖形的不穩定性，可能同時存在數個互相矛盾的 DAG。

可解決之辦法：

(1) 可放寬型一誤差之門檻，降低型二誤差的機率。

(2) 利用修正 p-value 或修正型一誤差α^*，提高條件獨立檢定之正確率。

(3) 距離遙遠之兩節點，其條件相關性之關係可忽略。

3-2 模擬變數，利用貝氏計分法建構基因網路

考慮兩變數 X、Y，若在不知道 X 與 Y 之分配的情形下，僅從資料上作統計分析，其因果模型：X→Y 或 Y→X，是等價的，我們無法從資料上去區分出兩圖形的差異。若一旦有變數分配的資訊，便可藉由貝氏計分法，確認出為何種因果關係。

而有些模型，即使在分配已知的情形下，亦會使 X→Y 與 Y→X 之因果關係為等價。如 $X \sim N(0,1)$ 、$Y = aX + b + \varepsilon$ ，$\varepsilon \sim N(0,1)$ 且 $X \perp\!\!\!\perp \varepsilon$。因此在上述關係中，其因果模型為 X→Y。而 X 亦可由 Y 表

示為：$X = cY + \varepsilon'$ 且 $Y \perp \varepsilon'$，其中 $c = \frac{a}{a^2+1}, \varepsilon' = X - \frac{a}{a^2+1}Y$（推導過程見附錄二），因此因果模型為 Y→X。欲使貝氏計分法區別出正確之因果關係，因此本節將不討論此類模型。

如 2.3-2 節定義，貝氏計分

$$S(G:D) = \log P(G \mid D) = \log \frac{P(D \mid G)P(G)}{P(D)} = \log P(D \mid G) + \log P(G) + C$$，

其中 G 為選定之基因網路圖形，D 為給定之資料，C 為一常數，與 G 的選取無關。由定義可知，貝氏計分可由兩部分所構成：$(1)\, P(D \mid G)(likelihood)$、$(2)\, P(G)(prior)$。本節欲模擬兩變數，利用貝氏計分法，判斷變數間之因果方向。模擬 model 如下：

X 為因(causality)，Y 為果(effect)，且 $X_i \sim Exp(\theta_1)$、$Y_i \sim Exp(\frac{\theta_2}{X_i})$，

$i = 1,...,n$，其中 $\theta_1 = 3, \theta_2 = 10$。其因果圖形為：X → Y。

考慮兩變數所可能造成的因果圖形(圖 3.4)，並計算其對應之貝氏計分。

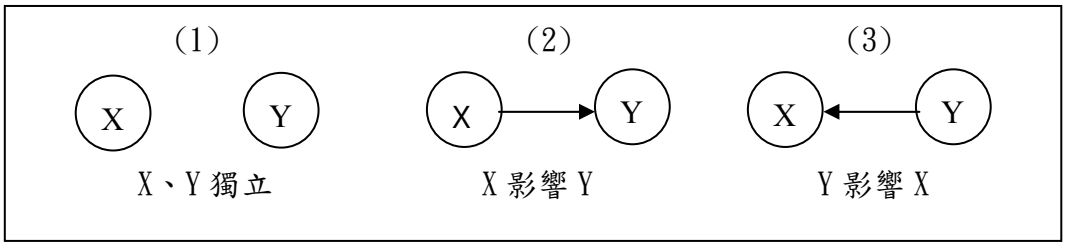


圖 3.4 三種因果關係圖

(1)假若 X、Y 之因果圖形 G_1 為： 

(X、Y 獨立)

則在 G_1 圖形下，設 $\begin{matrix} X_1,...,X_n \sim Exp(\theta_1), \theta_1 \sim Exp(\lambda_1) \\ Y_1,...,Y_n \sim Exp(\theta_2), \theta_2 \sim Exp(\lambda_2) \end{matrix}$，其中 λ_1, λ_2 為超參數

(hyper parameters)。故在此圖形下之 likelihood function 為

$$f_{G_1}(\tilde{x},\tilde{y} \mid \lambda_1,\lambda_2) = \iint f_{G1}(\tilde{x},\tilde{y} \mid \theta_1,\theta_2) \cdot \pi_{G_1}(\theta_1,\theta_2 \mid \lambda_1,\lambda_2) d\theta_1 d\theta_2$$

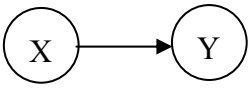
$$= \iint (\theta_1^n e^{-\theta_1 \sum x_i} \cdot \theta_2^n e^{-\theta_2 \sum y_i})(\lambda_1 e^{-\lambda_1 \theta_1} \cdot \lambda_2 e^{-\lambda_2 \theta_2}) d\theta_1 d\theta_2$$

$$= \lambda_1 \lambda_2 \frac{\Gamma(n+1)^2}{(\sum x_i + \lambda_1)^{n+1} (\sum y_i + \lambda_2)^{n+1}}$$

因為 λ_1, λ_2 為超參數(hyper parameters)，故利用 MLE 求其估計量，可得

$\hat{\lambda}_1 = \frac{1}{n}\sum x_i, \hat{\lambda}_2 = \frac{1}{n}\sum y_i$ ，則 G_1 的 likelihood function

$$f_{G_1}(\tilde{x},\tilde{y} \mid \hat{\lambda}_1,\hat{\lambda}_2) = \bar{x} \cdot \bar{y} \frac{\Gamma(n+1)^2}{(\sum x_i + \bar{x})^{n+!} (\sum y_i + \bar{y})^{n+!}}。 \quad (3.3)$$

(2)假若 X、Y 之因果圖形 G_2 為： 

(X 影響 Y)

則在 G_2 圖形下，設 $\begin{matrix} X_1,...,X_n \sim Exp(\theta_1), \theta_1 \sim Exp(\lambda_1) \\ Y_i \sim Exp(\frac{\theta_2}{X_i}), i = 1,...,n, \theta_2 \sim Exp(\lambda_2) \end{matrix}$，其中 λ_1, λ_2 為超

參數(hyper parameters)。故在此圖形下之的 likelihood function 為

$$f_{G2}(\tilde{x},\tilde{y} \mid \lambda_1,\lambda_2) = \iint f_{G2}(\tilde{x},\tilde{y} \mid \theta_1,\theta_2) \cdot \pi_{G2}(\theta_1,\theta_2 \mid \lambda_1,\lambda_2) d\theta_1 d\theta_2$$

$$= \iint (\theta_1^n e^{-\theta_1 \sum x_i} \cdot \theta_2^n \frac{1}{\prod_{i=1}^{n} x_i} e^{-\theta_2 \sum \frac{y_i}{x_i}})(\lambda_1 e^{-\lambda_1 \theta_1} \cdot \lambda_2 e^{-\lambda_2 \theta_2}) d\theta_1 d\theta_2$$

$$= \lambda_1\lambda_2 \cdot \frac{1}{\prod_{i=1}^{n} x_i} \cdot \frac{\Gamma(n+1)^2}{(\sum x_i + \lambda_1)^{n+1}(\sum \frac{y_i}{x_i} + \lambda_2)^{n+1}}$$

因為λ_1, λ_2為超參數(hyper parameters)，故利用 MLE 求其估計量，可得

$\hat{\lambda}_1 = \frac{1}{n}\sum x_i, \hat{\lambda}_2 = \frac{1}{n}\sum \frac{y_i}{x_i}$ ，則 G_2 的 likelihood function

$$f_{G2}(\tilde{x}, \tilde{y} \mid \hat{\lambda}_1, \hat{\lambda}_2) = \bar{x} \cdot (\frac{1}{n}\sum \frac{y_i}{x_i}) \frac{\Gamma(n+1)^2}{(\sum x_i + \bar{x})^{n+!}(\sum \frac{y_i}{x_i} + \frac{1}{n}\sum \frac{y_i}{x_i})^{n+!}} \text{。} \qquad (3.4)$$

(3)假若 X、Y 之因果圖形 G_3 為： X ← Y **(Y 影響 X)**

$$Y_1, ..., Y_n \sim Exp(\theta_1), \theta_1 \sim Exp(\lambda_1)$$

則在 G_3 圖形下，設 $X_i \sim Exp(\frac{\theta_2}{Y_i}), i = 1, ..., n, \theta_2 \sim Exp(\lambda_2)$ ，其中λ_1, λ_2為超參數(hyper parameters)。故在此圖形下之的 likelihood function 為

$$f_{G3}(\tilde{x}, \tilde{y} \mid \lambda_1, \lambda_2) = \iint f_{G3}(\tilde{x}, \tilde{y} \mid \theta_1, \theta_2) \cdot \pi_{G3}(\theta_1, \theta_2 \mid \lambda_1, \lambda_2) d\theta_1 d\theta_2$$

$$= \iint (\theta_1^n e^{-\theta_1 \sum y_i} \cdot \theta_2^n \frac{1}{\prod_{i=1}^{n} y_i} e^{-\theta_2 \sum \frac{x_i}{y_i}})(\lambda_1 e^{-\lambda_1 \theta_1} \cdot \lambda_2 e^{-\lambda_2 \theta_2}) d\theta_1 d\theta_2$$

$$= \lambda_1\lambda_2 \cdot \frac{1}{\prod_{i=1}^{n} y_i} \cdot \frac{\Gamma(n+1)^2}{(\sum y_i + \lambda_1)^{n+1}(\sum \frac{x_i}{y_i} + \lambda_2)^{n+1}}$$

因為λ_1, λ_2為超參數(hyper parameters)，故利用 MLE 求其估計量，可得

$\hat{\lambda}_1 = \frac{1}{n}\sum y_i, \hat{\lambda}_2 = \frac{1}{n}\sum \frac{x_i}{y_i}$ ，則 G_3 的 likelihood function

$$f_{G3}(\tilde{x},\tilde{y}\,|\,\hat{\lambda}_1,\hat{\lambda}_2)=\bar{y}\cdot(\frac{1}{n}\sum\frac{x_i}{y_i})\frac{\Gamma(n+1)^2}{(\sum y_i+\bar{y})^{n+!}(\sum\frac{x_i}{y_i}+\frac{1}{n}\sum\frac{x_i}{y_i})^{n+!}}\text{。}\qquad(3.5)$$

模擬資料：$X_i \sim Exp(\theta_1)$、$Y_i \sim Exp(\frac{\theta_2}{X_i})$，$i=1,...,100$，其中$\theta_1=3,\theta_2=10$。

將所生成資料代入上述各圖形所對應之 likelihood function $f(D\,|\,G)$（(3.3)~(3.5)）並取自然對數，記錄$\log f(D\,|\,G)$之值，比較其大小，重複模擬一百次，可得其平均數如下表：

表 3.11 各圖形之$\log f(D\,|\,G)$

	$G_1: X \perp Y$	$G_2: X \to Y$	$G_3: X \leftarrow Y$
$\log f(D\,|\,G)$ 最大出現次數	0	100	0
平均$\log f(D\,|\,G)$	251.236	308.601	175.951

由表 3.11 發現，正確的因果圖形 G_2，與所預期結果相同，其$\log f(D\,|\,G)$之值最大。

由定義可知，貝氏計分$S(G:D)=\log p(D|G)+\log p(G)+C$。因此，本節欲設定數組不同之圖形先驗機率 P(G)，並計算其貝氏計分(Score)。

表 3.12 各圖形之貝氏計分

圖形	P(G)	Log P(G)	log f(D|G)	Score	結論
G_1	0.33	-1.109	251.236	250.127	
G_2	0.33	-1.109	308.601	307.492	G_2：X→Y
G_3	0.33	-1.109	175.951	174.842	
G_1	0.50	-0.693	251.236	250.543	
G_2	0.25	-1.386	308.601	307.215	G_2：X→Y
G_3	0.25	-1.386	175.951	174.565	
G_1	0.80	-0.223	251.236	251.013	
G_2	0.10	-2.303	308.601	306.298	G_2：X→Y
G_3	0.10	-2.303	175.951	173.648	

由表 3.12 可知，不論在何種圖形 G 之先驗機率下，G_2均有最高之貝氏計分。因此由模擬結果，貝氏計分法可確認出正確之因果圖形。

此外，我們若考慮以不同的 model 來生成資料，且因果關係仍為 X→Y，但不改變 likelihood function，即亦繼續採用(3.3)～(3.5)之公式，探討以錯誤之 likelihood function，貝氏計分法是否仍可確認出正確之因果關係。其結果模擬如下：

表 3.13 不同模型之貝氏計分

Model: $X \sim U(1,5)$ $Y = 3X + \varepsilon, \varepsilon \sim N(0,1), X \perp \varepsilon$	$G_1 : X \perp Y$	$G_2 : X \rightarrow Y$	$G_3 : X \leftarrow Y$
$\log f(D \mid G)$ 最大出現次數	0	91	9
平均$\log f(D \mid G)$	-532.463	-523.883	-525.194

Model: $X \sim x_1^2$ $Y = X + \varepsilon, \varepsilon \sim N(5,1), X \perp \varepsilon$	$G_1 : X \perp Y$	$G_2 : X \rightarrow Y$	$G_3 : X \leftarrow Y$
$\log f(D \mid G)$ 最大出現次數	0	0	100
平均$\log f(D \mid G)$	-382.870	-855.521	-360.620

Model: $X \sim Weibull(3,0.5)$ $Y = 5 + X + \varepsilon, \varepsilon \sim N(0,1), X \perp \varepsilon$	$G_1 : X \perp Y$	$G_2 : X \rightarrow Y$	$G_3 : X \leftarrow Y$
$\log f(D \mid G)$ 最大出現次數	0	0	100
平均$\log f(D \mid G)$	-382.365	-493.411	-349.449

由模擬結果可發現，評估錯誤的 likelihood function，將使貝氏計分法的效能不彰，無法確認出正確的基因網路圖形。因此慎選各圖形之 likelihood function，為貝氏計分法不可忽略的環節。有關貝氏計分法更詳盡的模型討論，可參照交通大學統研所戴如美之碩士論文(2005)。

第四章　模型與(條件)獨立性之關係

4-1 不完全之條件獨立性檢定對模型之影響

由於樣本數的限制，本論文所探討變數間的條件獨立關係，均建構在給定一變數下，進行統計分析。因此對於需在給定兩變數，甚至給定更多變數的情形下，才能顯示出條件獨立關係之變數，本論文將不深入探討。

假若變數間存在需給定兩變數以上而呈現獨立之關係，我們亦以給定一變數下，進行條件獨立分析，會產生何種錯誤的因果圖形呢？錯誤的圖形對變數間之關係又會造成何種影響呢？

舉例說明，如圖 4.1，其真實模型之條件獨立關係為：$B \perp C \mid A$、$A \perp D \mid B,C$。若以 PC-Algorithm，且只檢驗給定一變數下之條件獨立關係，僅得：$B \perp C \mid A$，且 $D \notin S_{BC}$，並利用定向規則(R_3)，可得一 PDAG。

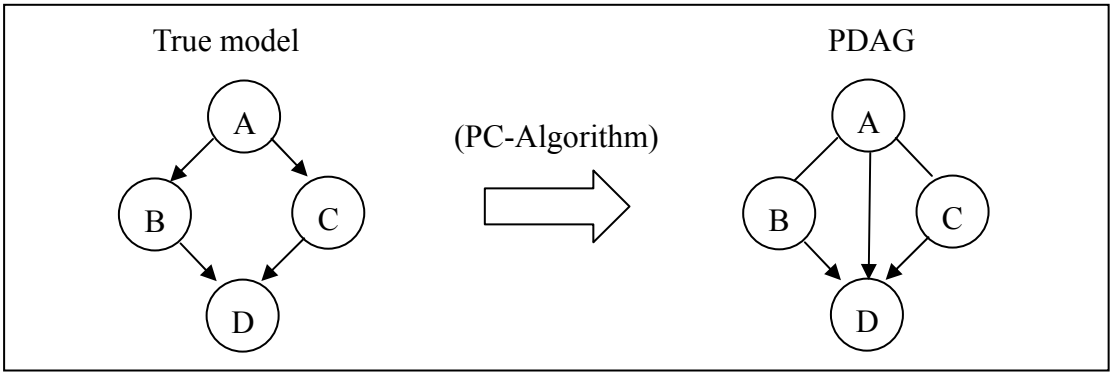


圖 4.1 不完全之條件獨立性檢定所建構之 PDAG(一)

再舉一例說明，如圖 4.2，其真實模型之條件獨立關係為：$A \perp B$、$A \perp D$、$C \perp D \mid B$、$A \perp E \mid C,D$、$B \perp E \mid C,D$。再以 PC-Algorithm，亦只檢驗給定一變數下之條件獨立關係，可得：$A \perp B$、$A \perp D$、$C \perp D \mid B$，且$C \notin S_{AB}$、$E \notin S_{CD}$，並利用定向規則(R_2)，可得一 PDAG。

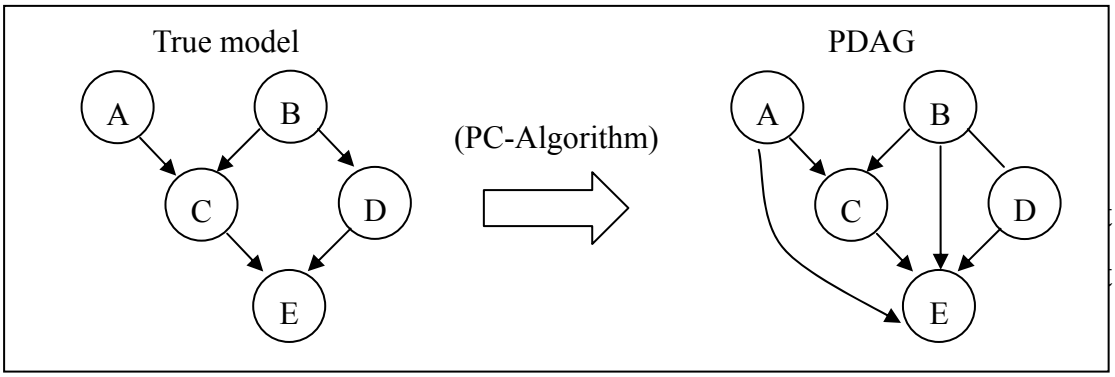


圖 4.2 不完全之條件獨立性檢定所建構之 PDAG(二)

由上述二例可發現，若變數間存在需給定兩變數之條件獨立關係，均以給定一變數，進行條件獨立性檢定，會使原本是"間接因果相關"之變數，變為"直接因果相關"，但該變數間仍有相同之因果方向。

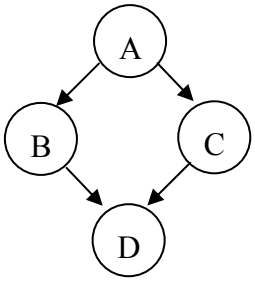
4-2 錯誤模型對(條件)獨立關係之影響

由第三章之模擬結果和文獻探討可知，在有限的資料下，因型二誤差β之影響，不易檢定出所有(條件)相關，故產生多餘的(條件)獨立關係，造成錯誤的基因網路圖形。而多餘的(條件)獨立關係，會使圖形中連接兩變數之線段消失，因此本節將討論圖形中因檢定而消失的線段，將會使(條件)獨立關係造成何種影響。

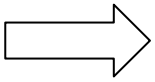
舉例說明，

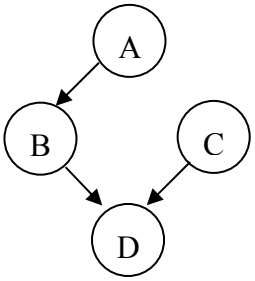
(1)

真實之基因網路



(除去 A-C 線段)





(條件)獨立關係：

$B \perp C \mid A$

$A \perp D \mid B,C$

(條件)獨立關係：

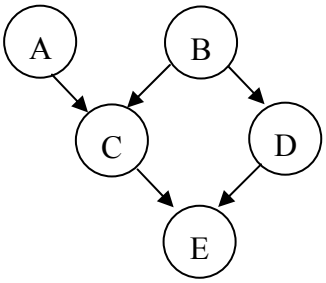
$A \perp C$

$B \perp C$

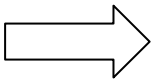
$A \perp D \mid B$

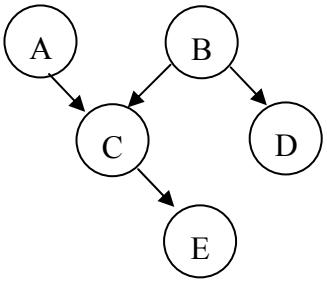
(2)

真實之基因網路



(除去 D-E 線段)





(條件)獨立關係：

$A \perp B$

$A \perp D$

$C \perp D \mid B$

$A \perp E \mid C,D$

$B \perp E \mid C,D$

(條件)獨立關係：

$A \perp B$

$A \perp D$

$C \perp D \mid B$

$A \perp E \mid C$

$B \perp E \mid C$

$D \perp E \mid B$

由上述兩例可發現，少一線段，會多一(條件)獨立關係(如上頁白底實線框框內之關係)，而真實模型原本之(條件)獨立關係則繼續傳遞至線段不全之模型，但其條件獨立關係部分有降階之現象，即原本給定一變數或兩變數之條件獨立關係，在線段不全之模型下，不需給定變數或給定一變數即呈(條件)獨立(如上頁灰底虛線框框內之關係)。

第五章　結論與展望

本論文利用 PC-Algorithm 和貝氏計分法所建構之基因網路，探討變數間之因果關係，由模擬結果可知，均可顯現出不錯之效能。而此兩法各有其優缺點，其比較如下：

PC-Algorithm

優點：對於母體分配不需做任何假設

缺點：由於樣本數之限制和誤差之影響，易造成偏大的型二誤差 β，產生多餘的(條件)獨立關係，且所建構之模型常為一 PDAG，即無法確認出變數間所有之因果方向，因此模型之正確率較低。

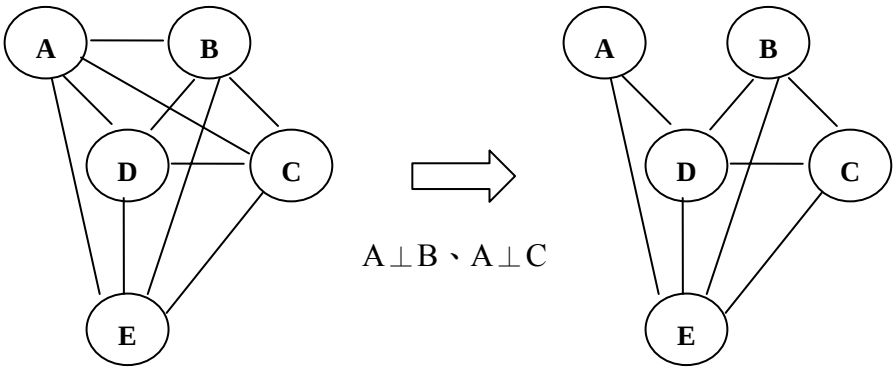
貝氏計分法

優點：確認出之基因網路通常為一 DAG，即可判斷出所有因果方向，且正確率高。

缺點：需有母體分配之資訊，不當的母體假設，會使模型產生嚴重的錯誤，且當變數個數增加，所涉及之模型個數呈指數倍成長，計算複雜，使此法效能低落。

由上述分析可發現，此兩法之優缺點，有互相補強之效果。因此在未來建構基因網路模型時，我們可考慮綜合兩法之特性，先以 PC-Algorithm 有效地縮減模型空間，對於所建構出之 PDAG，其因果方向未被確定之部分，再以貝氏計分法確認之。

附錄一



因此，

(1) B —— D —— E

(a) $B \perp D \mid E$ ➔ B —— E —— D

(b) $B \perp E \mid D$ ➔ B —— D —— E

} 矛盾

(1) B —— C —— E

(a) $B \perp E \mid C$ ➔ B —— C —— E

(b) $C \perp E \mid B$ ➔ C —— B —— E

} 矛盾

(1) C —— D —— E

(a) $C \perp D \mid E$ ➔ C —— E —— D

(b) $C \perp E \mid D$ ➔ C —— D —— E

} 矛盾

附錄二

$X \sim N(0,1)$ 、$Y = aX + b + \varepsilon$ ，$\varepsilon \sim N(0,1)$ 且 $X \perp \varepsilon$ ，a、b 為常數。

➔ $Y \sim N(b, a^2 + 1)$

➔ $\exists c \neq 0$ s.t. $X = X + cY - cY = cY + (X - cY)$

令 $\varepsilon' = (X - cY)$ 且 $Y \perp \varepsilon'$

因此 $Cov(Y, \varepsilon') = Cov(Y, X - cY)$

$$= Cov(Y, X) - c \cdot Cov(Y, Y)$$

$$= Cov(aX + b + \varepsilon, X) - c \cdot Var(Y)$$

$$= a \cdot Cov(X, X) + Cov(\varepsilon, X) - c \cdot Var(Y)$$

$$= a \cdot 1 + 0 - c \cdot (a^2 + 1)$$

$$= a - c(a^2 + 1)$$

$$= 0$$

➔ $c(a^2 + 1) = a$

➔ $c = \frac{a}{a^2 + 1}$

➔ $\varepsilon' = X - cY = X - \frac{a}{a^2 + 1} Y$

參考文獻

1. Constantin F. A. and I. Tsamardinos, (2003). Discovery of Causal Structure Using Causal Probabilistic Networks Induction. Retrieved August 27, 2004, from http://discover1.mc.vanderbilt.edu/discover/public/ml_tutorial_old/Presentation/part5.ppt.
2. Cooper, G. F., (1990). Computational complexity of probabilistic inference using Bayesian belief nework. *Artificial Intelligence* 42: 393-405.
3. Heckerman, D., (1996). A tutorial on learning with Bayesian networks, Microsoft Research tech. report, MSR-TR-95-06.
4. Kiiveri, H., T. P. Speed, and J. B. Carlin, (1984). Recursive causal models. *Journal of the Australian Mathematical Society* 36:30-52.
5. Lauritzen, S. L., (1982). *Lectures on Contingency Tables,* 2nd ed. Aalborg, Denmark: University of Aalborg Press.
6. Madsen, A. L., Lang, M., Kjærulff, U. B. and Jensen, F. (2003), The Hugin Tool for Learning Bayesian Networks. In Nielsen, T. D. and Zhang, N. L. (eds.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 7th European Conference, ECSQARU 2003*. Lecture Notes in Computer Science, pp. 594-605, Springer-Verlag Heidelberg.
7. Moral, S., (2003). Structure Learning: The PC Algorithm. Retrieved October 13, 2004, from http://www.dina.dk/phd/s/s6/learning3.pdf.
8. Neapolitan, R.E., (1990). *Probabilistic Reasoning in Expert Systems: Theory and Algorithms.* Wiley, New York.
9. Oliver, R.M. and Smith, J.Q., (1990). Influence Diagrams, Belief Nets, and Decision Analysis. John Wiley, New York.
10. Pearl, J., (1986). Fusion, propagation and structuring in belief networks. Artificial Intelligence 29, 241-288.
11. Pearl, J., (1995). Causal Diagrams for Empirical Research. Biometrika, 82, 669-709.
12. Pearl, J., (2000). CAUSALITY : Models, Reasoning, and Inference. University of California, Los Angeles.
13. Pe'er, D., Regev, A., Elidan, G., and Friedman, N., (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics 17 Suppl 1*: S215-S224.
14. Shachter, R.D., (1990), Special issue on influence diagrams. Networks: *An International Journal*, 20(5).

15. Shafer, G. and Pearl, J., (1990), *Readings in Uncertain Reasoning*. Morgan Kaufmann. San Mateo, CA.

16. Siegel, S., (1956). Nonparametric Statistics. McGraw-Hill. London.

17. Siegel, S. and Castellan, N. J., (1988). Nonparametric Statistics.

18. Spirtes, P., Glymour, C., and Scheines, R. (1993). Causation, Prediction, and Search. *Springer-Verlag*, New York.

19. Spirtes, P., C. Glymour, and R. Scheines. Constructing bayesian network models of gene expression networks from microarray data. In Proceedings of the Atlantic Symposium on Computational Biology. Genome Information Systems & Technology, 2000.

20. Verma, T., and J. Pearl, (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the 4th Workshop on Uncertainty in Artificial Intelligence* (Mountain View, CA), pp. 352-9. Reprinted in R. Shachter, T. S. Levitt, and L. N. Kanal (Eds.), *Uncertainty in Artificial Intelligence*, vol.4, pp. 69-76.Amsterdam: Elsevier.

21. Verma, T., and J. Pearl, (1990). Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence* (July, Cambridge, MA), pp. 220-7.Reprinted in P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer (Eds.). *Uncertainty in Artificial Intelligence*, vol.6, pp. 255-68.Amsterdam: Elsevier.

22. Verma, T., and J. Pearl, (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In D. Dubois, M. P. Wellman, B. D'Ambrosio, and P. Smets (Eds), Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence, pp. 323-30. Stanford, CA: Morgan Kaufmann.

23. Wermuth, N., and S. L. Lauritzen, (1983). Graphical and recursive midels for contingency tables. *Biometrika* 70:537-53.