変数間の寄与率に基づく因果関係から成る因果ダイアグラム

柴　田　淳　子

1．は　じ　め　に

因果関係は，「いくつかの事柄の関係において，一方が原因で他方が結果であるというつながりのあること。(大辞林)」とされている。因果関係の研究は古くから哲学をはじめ様々な分野において成されている。一般に，現象Xが現象Yの原因である場合，現象Yの方が現象Xより後に起こるとされている（時間的前後関係）が，これと因果関係とは必ずしも結びつくとは限らず，同時に起こる場合もあれば，現象Xの方が後に起こる場合もある。因果関係の決定に関して，時間的前後関係は重要な要因の一つではあるが，絶対的な条件であるとは言い難い。

因果推論では，原因と結果の決定すなわち因果の方向を取り扱う研究と，その方向は事前に決まっているという前提条件のもとで，その因果の大きさに焦点を当てた研究の2つに分かれる。本研究では，時間的前後関係が明らかでないような状況における前者の研究を対象とする。本論文において，我々は，複数の変数間の因果関係について，変数を頂点，その関係を辺としたグラフ理論におけるグラフ，つまり，因果ダイアグラムを用いて表現することを試みる。具体的には，Dempster によって提案された共分散選択を利用し，無向独立グラフを作成した後，これにより得られた無向独立グラフの辺の向き，つまり矢印の大きさを決定する方法を提案する。

2．因　果　関　係

竹内（1989）によると，因果関係の推定には次の5つの条件を考慮することが必要であるとされている。

（ｉ）時間的先行性

（ii）関連性の強さ

（iii）直接的関係

（iv）関連の一致性・普遍性

（ｖ）関連の整合性

ここでは，これまでに提案されている因果の方向を同定する方法について述べる。調査データを得る際に可能であれば，$X=x_0$ または $X=x_1$ を無作為割り付けし Y を観測することで，X から Y への因果の大きさを評価する。同様に，$Y=y_0$ または $Y=y_1$ を無作為割り付けし X を観測することで，Y から X への因果の大きさを評価した結果，両者を比較する方法が有効である。しかし，データ採取時の前提条件を全く同じものに揃える等の条件が必要である。

時刻 t と $t'(t'>t)$ の2時点において，データ $\{X_t, Y_t, X_{t'}, Y_{t'}\}$ が得られたとする。このとき，以下の2つのモデルにおいて係数を推定し，$\hat{\beta}_1$ と $\hat{\gamma}_1$ を比較することで因果の方向を決定することができる。

$$\begin{aligned} Y_{t'}-Y_t&=\beta_0+\beta_1 X_t+\varepsilon \\ X_{t'}-X_t&=\gamma_0+\gamma_1 Y_t+\varepsilon \end{aligned} \qquad (1)$$

ここで，$|\hat{\beta}_1|>|\hat{\gamma}_1|$ であれば，$X\rightarrow Y$ の因果がより強いと判断できるため，X が Y の原因であることを表している。

また，次のようなモデルも提案されている。

$$\begin{aligned} Y_{t'}&=\gamma Y_t+\beta_0+\beta_1 X_t+\varepsilon \\ X_{t'}&=\beta X_t+\gamma_0+\gamma_1 Y_t+\varepsilon \end{aligned} \qquad (2)$$

この場合の因果の方向も(1)式の場合と同様である。Wainer（1991）は，原因の変数群を停止した場合の結果における変数群の時間的変化に応じて，この2

つの式を使い分けることを提案している。

3．提　案　手　法

分析の対象となる変数 Y とそれを説明するための補助変数 $X_1, X_2, \cdots, X_n$ が表1に示すような形で観測されているとする。

表1：観測データ

Y	X_1	X_2	$\cdots$	X_i	$\cdots$	X_j	$\cdots$	X_n
Y_1	x_{11}	x_{12}	$\cdots$	x_{1i}	$\cdots$	x_{1j}	$\cdots$	x_{1n}
Y_2	x_{21}	x_{22}	$\cdots$	x_{2i}	$\cdots$	x_{2j}	$\cdots$	x_{2n}
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Y_N	x_{N1}	x_{N2}	$\cdots$	x_{Ni}	$\cdots$	x_{Nj}	$\cdots$	x_{Nn}

ここで，変数間に時間的な前後関係が存在する場合でも，因果関係と必ずしも結びつくとは限らないため，変数 Y も含めた（$n+1$）個の変数に関する因果関係の可視化を試みる。

本論文では，まず，Dempster によって提案された共分散選択を利用し，無向独立グラフを作成する。この手法は，変数間の仮説が少ない場合に有効な知見を与える手法である。そして，これにより得られた無向独立グラフの辺の向き，つまり矢印の大きさを3.2節に述べる方法により決定する。

3.1　共分散選択による無向独立グラフ

ここでは，Dempster（1972）によって提案された共分散選択について説明する。p 個の変数について，その相関行列を $R=|\rho_{ij}|$，その逆行列の要素を ρ^{ij} とすると，X_1 と X_2 の偏相関係数は以下の式で与えられる。

$$\rho_{ij \bullet rest} = -\frac{\rho^{ij}}{\sqrt{\rho^{ii}\rho^{jj}}} \qquad (3)$$

ただし，$rest$ は，変数 X_i と X_j 以外の変数を意味し，それらの変数の影響を取

り除くことを表している。 ここで，$\rho_{ij \bullet rest}=0$ のとき，変数 X_i と X_j は条件付独立の関係にあり，グラフにおけるその変数間の辺は存在しない。

共分散選択の手順は，偏相関係数の値が小さい順にそれを 0 と設定し，それによって表されるモデルの適合度により最適なグラフを選択する，というものである。

3.2 頂点間の因果関係

次に，共分散選択により作成した無向独立グラフにおいて，辺が存在する頂点間の関係を調べる。ここでは，そのような頂点に対応する変数を変数 X_i，変数 X_j とし，それらの観測データを用いて，どちらが一方の変数に対してどのくらい寄与しているかという観点から因果関係を決定することを試みる。

取り扱うデータが量的データである場合，既存研究に従った方法で因果関係を決定することが適切であると考えられる。しかし，質的データを取り扱う場合は，データの散らばりが図 1 の右図に示すような偏った散らばりになってしまい，量的データと同じ方法で因果関係を決定することが困難となることがある。

これを解決するために，本研究では，変数 X_i を T_i 個の水準に設定し，それに対する変数 X_j のばらつきを分析する要因実験の解析方法を適用する。まず，データを以下の手順に従って並び変える。

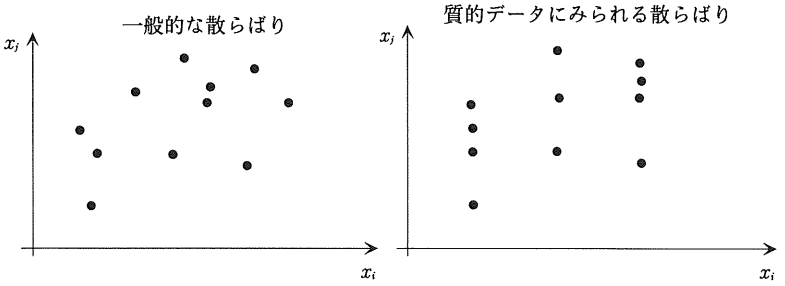


図 1 ：データの散らばり

表2：水準ごとに並び変えた観測データ

水準	※データ数は水準により一致しない			
A_1	t_{11}	t_{12}	$\cdots$	t_{1n_1}
A_2	t_{21}	t_{22}	$\cdots$	t_{2n_2}
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_{T_i}	t_{T_i1}	t_{T_i2}	$\cdots$	$t_{T_in_i}$

手順1：変数 X_i に対する変数 X_j のデータを1組 $(x_{1i}, x_{1j})(i, j=1, 2, \cdots, N)$ として，変数 X_i を昇順（もしくは降順）にソーティングする。そして，重複しない変数 X_i の個数を水準 T_i 個と設定する。

手順2：各水準 T_i に対するデータ t_{ij} を以下のように，1組 $(x_{1i}, x_{1j})(i, j=1, 2, \cdots, N)$ のデータに対して2つのデータを設定する。（表2）

$$t_{i1}=\alpha+\beta x_i-\bar{x}_j \tag{4}$$

$$t_{i2}=\alpha+\beta x_i-x_j \tag{5}$$

これにより，データ数は $2N$ 個となる。ただし，パラメータは以下の通りである。

$$\beta=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sum_i(x_i-\bar{x})^2} \tag{6}$$

$$\alpha=\bar{y}-\beta\bar{x}_j \tag{7}$$

$$\bar{x}_j=\sum_j x_j/N \tag{8}$$

このとき，分散分析は表3に示す通りである。

表3：分散分析

要因	平方和 s	自由度	分散
水準間 A	S_A	f_A	$V_A=S_A/f_A$
誤差 e	S_e	f_e	$V_e=S_e/f_e$
合計 T	S_T	f_T	

ただし，水準ごとのデータ数は異なるため，それぞれの計算式は以下のようになる。

$$S_T=\sum_i\sum_j t_{ij}^2-CT \tag{9}$$

ここで，

$$CT=\frac{\left\{\sum_i\sum_j t_{ij}\right\}^2}{\sum_i n_i} \tag{10}$$

とすると，水準間の平方和は以下の式となる。

$$S_A=\frac{\left\{\sum_{i\in A_1}\sum_j t_{ij}\right\}^2}{n_1}+\cdots+\frac{\left\{\sum_{i\in A_i}\sum_j t_{ij}\right\}^2}{n_i}-CT \tag{11}$$

また，自由度は以下の式で表わされる。

$$\begin{aligned} f_A&=n_i-1(=\text{水準数}-1)\\ f_T&=N-1(=\text{総データ数}-1)\end{aligned} \tag{12}$$

このとき，要因によるばらつきが，全体の変動のどのくらいを占めているかを調べるために，以下の式により寄与率を計算する。

$$\rho_A=\frac{S_A-f_A MS_e}{S_T}\times 100 \tag{13}$$

これにより，変数 X_j の変動に変数 X_i がどのくらい寄与しているかが求まる。次に，変数 X_j についても同様に T_j 個の水準に設定し，寄与率 ρ_A' を求める。そして，両方の寄与率の値を比較することで，変数間の因果関係が明らかになる。つまり，$\rho_A>\rho_A'$ のとき，$X_i\rightarrow X_j$ の因果がより強いと判断し，X_i が X_j の原因であるとする。

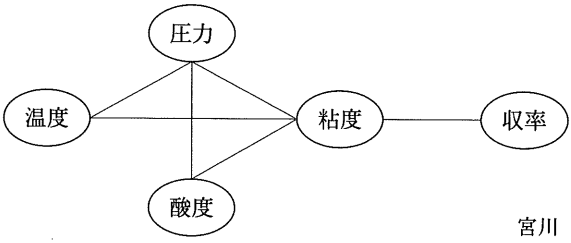
4．数　値　実　験

ここでは，表4に示す圓川（1988）の科学工程の収率と製造条件に関するデータを用いて数値実験を行う。また，このデータに対して共分散選択を行った結果は図2に示す通りである。

表４：数値実験に使用したデータ

No.	圧力	温度	酸度	粘度	収率
1	14.5	87.6	7.5	6.2	30.4
2	17.1	89.3	6.9	5.5	26.5
3	16.5	92.3	7.2	5.7	29.2
4	15.5	89.2	7.4	6.1	29.5
5	16.6	87	6.5	5	25.9
6	18.8	91.6	8.2	5.9	29.6
7	19.1	90	7.3	5	26.2
8	17.5	91.5	7.8	5.7	28.1
9	14.6	89.7	7	6.4	31.1
10	16.1	90.5	6.7	5.2	26.9

圓川（1988）



宮川（1997）

図２：無向独立グラフ

提案方法により得られた因果ダイアグラムを図３に示す。矢印の横に示した数値は，変数間の寄与率である。

図３において，「圧力」と「温度」，「圧力」と「酸度」に関しては，変数間の寄与率が同じ値であるため，因果の大きさは同じであると考えられる。また，それ以外の変数間については，寄与率のより大きい方の値が示す矢印の因果が大きいため，矢印は太線で表したものになると考えられる。

変数間の寄与率に基づく因果関係から成る因果ダイアグラム

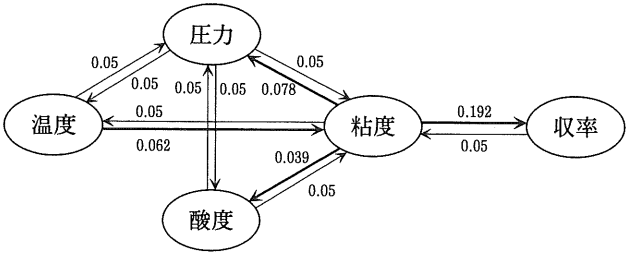


図3：数値実験により得られた因果ダイアグラム

5．おわりに

本論文では，複数の変数間の因果関係を因果ダイアグラムによって表現することを試みた。具体的には，まず，共分散選択に基づいた無向独立グラフを作成した後，これにより得られた無向独立グラフの辺に対して矢印の大きさを決定する方法を提案した。さらに，数値実験によりその有効性を示した。

参考文献

[1] 甘利俊一，狩野裕，佐藤俊哉，松山裕，竹内啓，石黒真木夫．(2002)．多変量解析の展開．岩波書店．

[2] 竹内啓．(1986)．因果関係と統計的方法．行動計量学，vol. 14, No. 1, pp. 85-90.

[3] 竹内啓．(1989)．統計学辞典．東洋経済新報社．

[4] 谷津進．(2001)．技術力を高める品質管理技法．朝倉書店．

[5] Dempster, A. P. (1972). Covariance selection. *Biometorics*, vol. 28, pp. 157-175.

[6] Wainer, H. (1991). Adjusting for Differential Base Rates: Lord's Paradox Again. *Psychological Bulletin,* vol. 109. No. 1. pp. 147-151.

[7] 圓川隆夫．(1988)．多変量のデータ解析．朝倉書店．

[8] 宮川雅巳．(1997)．グラフィカルモデリング．朝倉書店．