グラフィカルモデリングによる因果推定の研究

M2005MM026 榊原 浩晃
指導教員　松田 眞一

1　はじめに

今日，グラフィカルモデリングを用いて解析を行ったという研究内容が多数報告されている。しかし，いずれの報告においても解析結果は解析者の経験則によって導かれ完結されている。これはグラフィカルモデリングの特性上，対話的処理を終了させる基準が存在しないためである。このため，真のモデルを導きだすための解析が，解析者の認識力に依存した結果しか導き出すことが出来ず，解析者によって解析結果は相違する。そもそも真のモデルを探索する解析であるため，解析者の判断はできる限り最小限にしなければならないといえる。

よって，数値計算のみでより真のグラフに近づける基準が必要となるが，共分散選択 (Dempster[1]) を行うことで導き出されるグラフィカルモデリングの結果だけでは因果の推定は不可能であることが分かる。ここで，統計的に因果の推定が可能となる SGS アルゴリズム (Spirtes，Glymour，and Scheines[5]) を用いることで統計的な因果推定が可能となる事が知られているが，理論上の話であるため実際のデータを用いて解析を行う事は難しい。本論文では，この 2 つの手法の解析結果が最良となる基準を明らかにする。

2　グラフィカルモデリング

2.1　グラフィカルモデリングとは

グラフィカルモデリングとは，事前にははっきりしていない因果関係や変数の絡み具合をデータに基づいて探索的にモデル化して，その妥当性を検証することのできる方法論である。それは，すべての変数間を線で結んだ完全無向グラフ（矢線の無いグラフ）を作成し，変数間の関係が独立であれば線を切るといった操作を繰り返すことで多変量データをモデリングし，本来のモデルを探索的に推論するものなどが提案されている。

グラフィカルモデリングの手順を簡潔にまとめたものを以下に示す。

グラフィカルモデリングの手順

1. 線断基準として，データの偏相関値が一番低いものを候補として出す。
2. 手順 1 の候補の偏相関値を 0 とする。
3. 逸脱度・P 値などが許容範囲であれば反復アルゴリズム (wermuth[7]) を用いて手順 1 を繰り返す。

しかし，今日グラフィカルモデリングを用いて解析を行ったという発表には，どのような基準で解析を行ったかが記されていない。それはほとんどの解析者は，統計的根拠を持たず，それぞれの経験則のみで解析を進めているからである。本論文ではグラフィカルモデリングにおいて大きな問題点は 3 つあると考えている。

1. **打ち切り問題**
   グラフィカルモデリングの終了条件をさす。P 値の値，次に線断する候補の偏相関係数の値，相関係数・偏相関係数の跳ね上がり値などから処理を終了するための基準が明確にされていない。
2. **合流問題**
   グラフィカルモデリングは真のモデルで存在する矢線以外に，矢線が合流をしている説明変数同士に生まれる線も導き出してしまう。
3. **打ち消し合い問題**
   合流のための線が生じるため，合流のための線と本来そこにあるべき矢線の大きさがほぼ同じであり，かつ符号が逆の場合解析中に誤って線断してしまうことが多々ある。

2.2　シミュレーションの説明

多くの書籍 (例えば [2][4]) や論文で基準とされている $p = 0.5$ は正しいかどうかを明らかにするために，打ち切り基準となるフルモデルとの適合度検定の P 値を動かしてシミュレーションを行う。そうすることで，P 値によって推定されるモデルはどのように変化をするかを見ることができる。また，それと同時に，相関係数の跳ね上がり係数値や，線断基準の値も監視しながらシミュレーションを行うことにより，P 値以外の打きり基準を提案することができる。本論文のシミュレーションでは非巡回的なモデルのみを扱い，末端となる頂点は 1 つだけとする。また，ランダムに真のモデルを作成する際，与えるパラメータは頂点の数のみとする。このとき，真のモデルの矢線の数は

$$\frac{n(n-1)}{4} + 1 \tag{1}$$

を計算して小数点を切り上げた値とする。ここで，頂点の数は n，サンプル数 N とする。

2.3　シミュレーションの結果

図 1 は $n = 4, 6, 8$ と，$N = 100, 500, 1000$ の時のそれぞれの P 値との関係グラフである。グラフからも分かるように，サンプル数や頂点数によってグラフの形状が大きく変わってきている。よってこれまでのように $p = 0.5$ と決めてしまうのは必ずしも正しくないことが分かり，サンプル数・頂点数に合った基準を決める必要があるといえる。この他にも相関係数・偏相関係数の跳ね上がり値や線断候補の偏相関係数値を調べたところ，相関係数の跳ね上がり値で終了となることはほとんど存

在しないため，表 1 には P 値・偏相関係数の跳ね上がり値・線断候補値のみ載せる。表中で数値が書かれていない箇所は，その他の基準によって決められてしまうため基準値が存在しないことを示す。

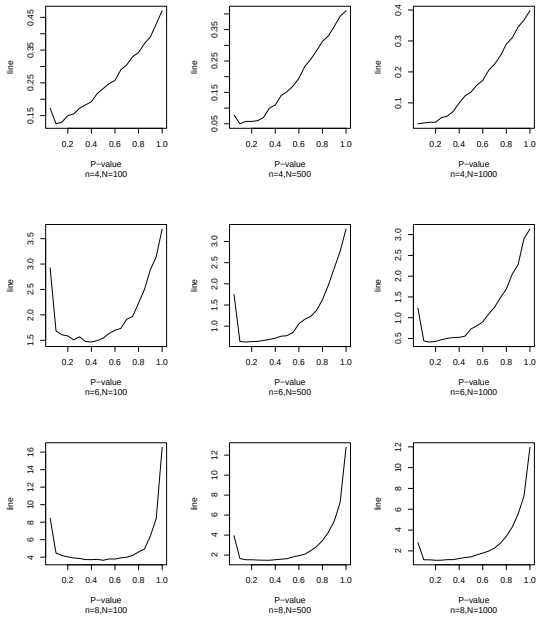

図 1　P値との関係

		P 値	偏相関係数	線断候補値
n=4	N=100	0.1	-	-
	N=300	0.1	-	-
	N=500	0.1	-	-
	N=700	0.1	-	-
	N=1000	0.1	-	-
n=5	N=100	0.15	-	-
	N=300	0.1	0.05	0.115
	N=500	0.1	0.045	0.12
	N=700	0.1	-	0.095
	N=1000	0.1	-	0.095
n=6	N=100	0.4	0.035	0.155
	N=300	0.3	0.025	0.11
	N=500	0.15	0.025	0.09
	N=700	0.2	0.025	0.075
	N=1000	0.15	0.02	0.065
n=7	N=100	0.45	0.04	0.11
	N=300	0.35	0.035	0.09
	N=500	0.2	0.03	0.08
	N=700	0.25	0.025	0.07
	N=1000	0.15	0.025	0.06
n=8	N=100	0.5	0.045	0.095
	N=300	0.35	0.035	0.085
	N=500	0.35	0.03	0.075
	N=700	0.25	0.03	0.07
	N=1000	0.2	0.025	0.06

表 1　打ち切り基準一覧表

3　SGS アルゴリズム

3.1　SGS アルゴリズムとは

グラフィカルモデリングは対象となる変数 X_i，X_j が独立かどうかのチェックは，この 2 変数以外の変数を条件付にする事で決めていた。つまり，全変数集合を

$V = \{X_1, X_2, \cdots, X_p\}$

とする場合，X_i，X_j の独立か否かについては

$I_{\{i,j\}} = \{X_1, \cdots X_{i-1}, X_{i+1}, \cdots, X_{j-1}, X_{j+1}, \cdots, X_p\}$

についての条件を付けて独立性を調べた。

しかし，これでは因果合流のための線が残ってしまうため独立グラフから因果の向きを考慮した有効グラフを推測する時に惑わされてしまう。グラフィカル連鎖モデリングではある程度考慮はされているが，同じ群内での因果合流には対応していない。また，実際の解析では常に必ずしも正しく群分けをする事ができないことが多いため，誤った因果合流を作成してしまうことも少なくない。そこですべての因果合流を発見することのできる SGS アルゴリズムが考えられた。この手法は理論上では真のモデルにある程度近づくことができる。しかし，実際のデータを用いて解析を行うと，揺らぎの関係で誤った推定を行ってしまうことがある。ここで，SGS アルゴリズムにおける問題点を挙げる。

1. **SGS アルゴリズムにおける線断問題**

   真のデータでは条件付独立となる場合の値は 0 である。しかし，実際のデータを用いると 0 より若干大きくなってしまう。よって，解析を行う上で条件付独立とみなす偏相関係数の値の範囲が重要となる。

2. **打ち消し合い問題**

   図 2 のようなグラフの場合グラフィカルモデリングの時は因果合流の時に生じる打ち消し合いである A → C の線が消えてしまうという問題だけであったが，SGS アルゴリズムにおける打ち消し合い問題はこの因果合流だけでない。A → B への直接的な大きさと直接効果以外の A → B への有向道の大きさの符号が逆の関係となり，さらに絶対の差が ϵ を下回ってしまうと A → B が消えてしまう現象が起こる。これは，SGS アルゴリズムの手順 4 で使われているオリエンテーションルール (Verma and Pearl[6]) においても考慮すべき点である。

SGS アルゴリズムの手順は以下に示す。(宮川 [3] 参照)

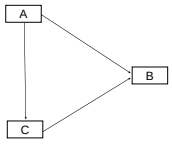

図 2　SGS による打ち消し合い問題

SGS アルゴリズム手順

1. 頂点集合が V である完全無向グラフ H を初期解として設定する。
2. V から任意の変数対 (X_i, X_j) が，$V \setminus \{X_i, X_j\}$ のある部分集合 S（空集合のこともある）を与えたとき条件付き独立であれば，完全グラフ H より X_i と X_j の間の辺を除去する。この結果，得られた無向グラフを K とする。
3. K において，X_i—X_k—X_j という構造（X_i と X_j は隣接していない）があるとき，X_k を含む頂点集合 S^* を与えたとき条件付き独立となるような S^* が存在しないとき，$X_i \rightarrow X_k \leftarrow X_j$ という矢線をつける。
4. K にいくつかの矢線が加わったグラフにおいて，$X_i \rightarrow X_k$—X_j という構造があり，X_i と X_j が隣接していないならば $X_k \rightarrow X_j$ と矢線をつける。
5. K にいくつかの矢線が加わったグラフにおいて，X_i から X_j に有向道があり，かつ，X_i と X_j の間に無向の辺があれば，その辺に $X_i \rightarrow X_j$ と矢線をつける。
6. この手順 4 と 5 を，矢線をつける辺がなくなるまで続ける。

3.2　シミュレーションの説明

真のモデルは 2.2 節で示した方法で作成する。線断基準となる α を 0 から 0.1 まで 0.02 間隔で 50 通りに分け，さらにそれぞれ 100 回ずつシミュレーションを行う。そこで，誤って真のモデルには存在する線を消してしまった時の線の数と，線を切ることができなかった時の数を求めることで，一番真のモデルに近い線断基準を導き出す。

この解析を頂点が 4〜8 つでサンプル数が 100 個，300 個，500 個，700 個，1000 個のすべてのパーターンにおいて行い，グラフィカルモデリングの時と同様，頂点数やサンプル数によってどのように結果が変わるかを調べる。

3.3　シミュレーションの結果

図 3 に示したように，サンプル数が多いほうが若干線断基準となる偏相関係数値を 0.01 ほど下げたほうが良いという結果となった。頂点数に関しては大きく差があり，頂点数が多ければ多いほど線断基準となる偏相関係数値は小さいほうが良いという結果となった。詳しい値は表 2 にある。

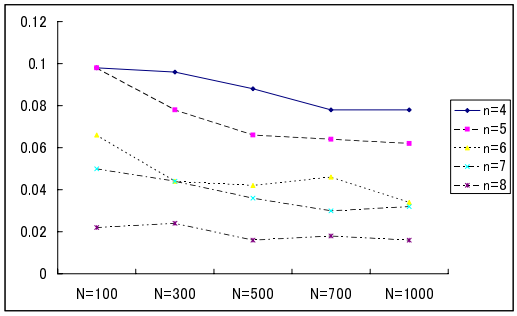

図 3　SGS アルゴリズムによる打ち切り基準プロット図

	n=4	n=5	n=6	n=7	n=8
N=100	0.098	0.098	0.066	0.050	0.022
N=300	0.096	0.078	0.044	0.044	0.024
N=500	0.088	0.066	0.042	0.036	0.016
N=700	0.078	0.064	0.046	0.030	0.018
N=1000	0.078	0.062	0.034	0.032	0.016

表 2　SGS アルゴリズムによる打ち切り基準一覧表

4　グラフィカルモデリングと SGS アルゴリズムの比較

これまでの結果より，グラフィカルモデリングと SGS アルゴリズムにおいて最良グラフのためのパラメータ値がわかった。しかし，２つの手法で求められたグラフは必ずしも同じ結果を導き出さないため最良グラフをそれぞれ作成するだけでは真のモデル推定ができたとはいえない。4 章では，求められたグラフを用いてその後どのようにグラフの推定を行うかを考察する。

4.1　比較方法

グラフィカルモデリングと SGS アルゴリズムの比較を行うために，真のグラフより作成した同じデータを用いてそれぞれ解析を行う。本章では $n = 7, N = 1000$ とし，設定するパラメータは表 1 と表 2 の最良基準を使った。また，SGS アルゴリズムにおける線断基準は 0.05 の場合も解析を行う。また，シミュレーション回数は 500 回とする。得られた結果より以下の６項目を数え上げる。

- 真のモデル以外の箇所に残ってしまった線の数（真のモデル以外）
- 真のグラフに矢線が存在しない場所に生まれた合流のための線の誤り数（合流１）
- 真のグラフに矢線が存在する場所に生まれた合流のための線の誤り数（合流２）
- 真のグラフに矢線が存在する箇所に対して消してしまった線の数（パス）

- 真のグラフに存在する矢線に対して，SGS には線があるが SGM には線が無い数（比較１）
- 真のグラフに存在する矢線に対して，SGM には線があるが SGS には線が無い数（比較２）

ここで，“誤りの数”と表記した理由は SGS アルゴリズムは合流のための線は理論上生まれないということを踏まえて，グラフィカルモデリングでの真のモデルはモラルグラフとし，SGS アルゴリズムでは真のモデルの矢線を線に変えた無向グラフを用いるからである。

4.2 比較結果

比較の結果を表 3 と表 4 に示す。図 4 と図 5 は縦軸がモデルの数であり、横軸は矢線が正しく推定できている率である。

このシミュレーションの頂点数は 7 であるため，真のグラフで作成される矢線の数は 12 本である。これを 500 回繰り返しているため矢線だけでも総数は 6000 本存在する。表 3 より，線断基準が 0.016 の場合の SGS アルゴリズムにおいて誤って消してしまう本数は $779 + 504 = 1283$(本) となり，グラフィカルモデリングの 152 本と比べ SGS アルゴリズムは消しすぎてしまう傾向がある。よって、SGS アルゴリズムによって求められた独立グラフを基に真のグラフを推定するのは非常に難しい。グラフィカルモデリングの場合は『合流２』の誤り数が 152 本と少なく，『パス』の誤り数は 0 となり完全に推定できている。さらに，表 4 より，SGS アルゴリズムで求められる線のほとんどがグラフィカルモデリングでカバーしていることが分かる。このことから、無向グラフだけの解析では SGS アルゴリズムを行うよりグラフィカルモデリングを行えば良いことが分かった。

ここで、SGS アルゴリズムを用いて矢線を推定した比較を行う。図 4 と図 5 を比べ、基準が 0.05 の方が正しく矢線を推定できている率が高くなっていることがわかる。これは、『真のモデル以外』の項目で 0.05 の場合は 18 本,『合流１』の数は 27 本と誤って推定する数が最良基準の 0.016 の時より数が少ないため矢線を正しく推定できた。

	真のモデル以外	合流１	合流２	パス
SGM	196	47	152	0
SGS(0.016)	96	134	779	504
SGS(0.05)	18	27	1009	650

表 3　比較の結果１

	比較１	比較２
SGS(0.016)	20	1151
SGS(0.05)	4	1511

表 4　比較の結果２

5 おわりに

今後の課題として，まず 4 節で示したグラフィカルモデリングと SGS アルゴリズムの比較において，本研究では SGS アルゴリズムの基準値を 0.05 とおいて話を進めたが，これもシミュレーションを行い，最良値を見つけることでより真のグラフに近い因果推定を行うことができるようになる。この場合における最良値というのは，SGS アルゴリズムによって作成される有向グラフと真のグラフにおいて，ただ正しく推定される数が最大となる基準を求めるというものではなく，間違えて推定してしまう数が極力少なく，かつある程度正しい矢線が推定できる基準である値を指す。その理由として，SGS アルゴリズムにおける無向グラフはグラフィカルモデリングで大半をカバーすることができるため，より信頼のできる有向グラフを作成することが一番最良の手段だと考えられるからである。

また，本研究では，真のグラフの矢線の数は固定したものを採用している。これに関しても矢線の数によってどのように結果が変わってくるかを明らかにすることで，より真のグラフに近いモデルを推定することが可能となる。

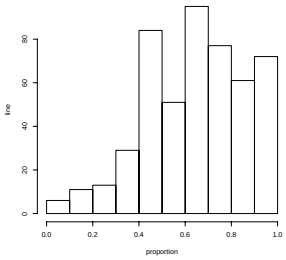

図 4　線断基準　0.016

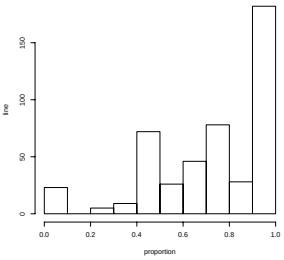

図 5　線断基準　0.05

参考文献

[1] Dempster , A.P.(1972). Covariance selection. Biometrics, **28**,157-175.

[2] 宮川雅巳 (1997)：グラフィカルモデリング, 朝倉書店.

[3] 宮川雅巳 (2004)：統計的因果推論, 朝倉書店.

[4] 日本品質管理学会 (1999)：グラフィカルモデリングの実際, 日科技連出版社.

[5] Spirtes, P., Glymour , C.and Scheines, R.(1990): Causality from probability. Proc. Conf. on Advanced Computing for the Social Sciences.

[6] Verma, T. and Pearl, J.(1992)：An algorithm for deciding if a set of observed independence has a causal explanation., Proc.8th Conf. on Uncertainty in AI: Stanford, Morgan Kaufmann, 323-330.

[7] Wermuth, N. and Scheidt, E.(1977) : Fitting a covariance selection to a matrix. Algorithm AS 105. Appl. Statist., **26**, 88-92.