

SGS アルゴリズムとグラフィカルモデリングに関する一考察

野口 良輔* 松田 眞一†

E-Mail: matsu@nanzan-u.ac.jp

SGS アルゴリズムは因果分析において有向線を導き出すことのできる重要な統計的方法論である。その理論の R への実装に関しては榊原 (2007) が行っている。本論文では、同じく榊原 (2007) が実装したグラフィカルモデリングの R 関数とともに EXCEL 上での作図マクロを実現し、それに基づいて 2 つの方法を用いた因果分析に関して考察する。

1 はじめに

データとして与えられた変数間の関係を調べる因果分析において、因果の向きが完全に分かっていてそのままパス解析が行えることは稀である。因果の向きが全く分からない、または一部は分かっている場合に因果関係の推定を行う方法にグラフィカルモデリングがある。グラフィカルモデリングでは偏相関関係があると思われる変数間を無向の線で結ぶ。一方、数学的独立性の観点から有向線を導き出す方法に SGS アルゴリズムがある。SGS アルゴリズムは数学的な方法であるため計算機上で実現はほとんどなく、そのためにこの 2 つの方法に関して比較した文献は見当たらない。

本論文では SGS アルゴリズムとグラフィカルモデリングの比較を行うために作図マクロの作成を行い、それに基づいた実データでの検討を行う。

2 用語と手順

後での議論のために SGS アルゴリズムとグラフィカルモデリングおよびパス解析に関する必要な用語を説明する。(宮川 [1], 宮川 [2] 参照。)

まず、因果の向きが完全に分かっていて変数間を有向線で結んだ図をパスダイアグラムという。パスダイアグラムができていればパス解析により、パス係数の推定が行えて変数間の関係をつぶさに見ることができる。

それに対して変数間の影響を他の変数の影響を除いた偏相関係数で捉え、それが無相関でない変数同士を無向の線で結んだグラフを無向独立グラフという。

グラフィカルモデリング（以下では GM と略することがある）とはデータから変数間の偏相関係数を求め、全部の変数間に線がある完全グラフから線を切断してこの無向独立グラフを得るための方法論である。具体的な手順は宮川 [1] を参照されたい。

一方、有向線を含むグラフを推定する場合、パスの経路が巡回していないことが重要であり、それを非巡回的と呼び、すべての線の方向を決めることができるグラフを非巡回的有向グラフと呼ぶ。さらに、各変数が上流から分布が決まっており、下流の変数が上流の変数を与えたときの条件付き分布で表現できるとき、そのグラフを非巡回的有向独立グラフと呼ぶ。非巡回的有向独立グラフは統計的に導かれるものであるため因果関係は推

* 南山大学大学院理工学研究科システム数理専攻

† 南山大学理工学部システム数理学科

定に過ぎず、そのままパスダイアグラムとなるわけではないことを注意されたい。因果関係については他の情報から推測することになり、数学的に同値な非巡回的有向独立グラフのうち正しいと思われるパスダイアグラムを導くことになる。

SGS (Sprites, Glymour and Scheines) アルゴリズム (以下では単に SGS と略すことがある) とは数学的な条件付き独立性から変数間の関係を調べて導出する方法論である。結果として、有向線と無向線の混ざった部分的矢線グラフが出力される。本論文での具体的な手順は以下の通りである。(宮川 [2] 参照)

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 からの任意の変数対 (X_i, X_j) が、 $V \setminus (X_i, X_j)$ のある部分集合 S (空集合の場合もある) を与えたときに条件付き独立であれば、 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 X_i と X_j が隣接しない道があるとき、 X_k を含む変数集合 S^* で、 X_i と X_j が S^* を与えたとき条件付き独立になるようなものが存在しないならば、 $X_i \rightarrow X_k \leftarrow X_j$ という矢印をつける。この操作を終えた部分的矢線グラフを P とする。
4. P において $X_i \rightarrow X_k - X_j$ という道があり、 X_i と X_j が隣接していないならば、 $X_k \rightarrow X_j$ と矢印をつけて P を更新する。
5. P 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 $X_i \rightarrow X_j$ と矢線をつけて P を更新する。
6. 手順 4, 5 を矢線をつける辺がなくなるまで続ける。

手順 4, 5 ではオリエンテーションルールと呼ばれる考え方が使われている。

なお、宮川 [2] にあるようにオリジナルの SGS アルゴリズムでは、手順 3 での変数に順序をつけて検討することになっているが、榊原 [3] は K から順次矢線に変更していくのではなく、該当する対について一度に変更する仕様を用いている。

3 解析支援ソフトウェアと仕様

グラフィカルモデリングと SGS アルゴリズムの解析支援ソフトウェアを EXCEL のマクロで実現した。実際のグラフィカルモデリングと SGS アルゴリズムの計算は榊原 [3] によって作製された R の関数 `sgm`, `sgs` を用いた。谷口 [5] がグラフィカルモデリングの解析支援ソフトウェアを作成しているが、それは使用せずに、浅井 [4] のパス解析支援ツールを改良することで 2 つの解析支援ソフトウェア仕様を近くし、グラフィカルモデリングと SGS アルゴリズムの解析からパス解析へと続けて解析を行う際、ユーザに違和感なく使用できるようにした。

グラフィカルモデリング、SGS アルゴリズムの支援ソフトウェアは、2 つの出力されたグラフを比較する際、結果比較の見安さからそれぞれ別の Book になっている。

以下にそれぞれの Book における大まかなソフトウェアの実行プロセスを 8 つに別けて説明する。

1. EXCEL で R の実行パスをユーザに指定させる。
2. テキストファイルのデータを「GM SGS 支援ツール」フォルダ内の「●table」フォルダの中に格納後にデータのファイルのパスを保存する。
3. R の実行命令文を VB スクリプトによって作成し、バッチ処理として R を実行し、計算を行う。
4. R の計算結果を R によってテキストファイルに保存する。
5. 保存されたテキストファイルの有無を EXCEL によって確認させる。
6. EXCEL 上に計算結果を出力させるためワークシートの初期化を行う。
7. EXCEL で R の計算結果のテキストファイルを読み込む。
8. EXCEL でグラフを作成する。

実際の解析支援ソフトウェアによる 2 つのブックの出力を並べた例は図 1 となる。2 つの Book での変数の座標は同期を取ることができるよう工夫されている。

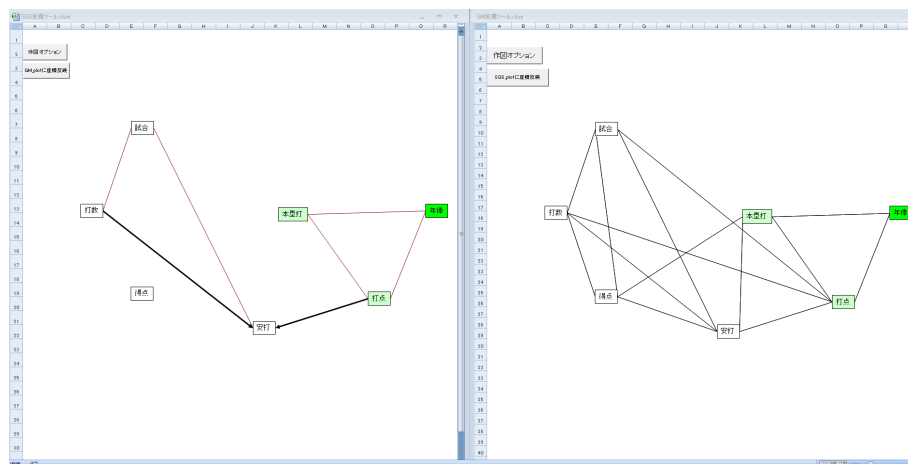


図 1: 2 つの作成支援ツールで作成されたグラフ例（左：SGS，右：GM）

4 データ解析例

政府統計から持ってきた 2004 年の都道府県別の月消費支出とその関連変数を用いて解析を行った。目的変数を月消費支出とにおいて、食費、住居費、年間収入、貯蓄現在高、負債現在高がどのように関わっているかを考える。

グラフィカルモデリング、SGS アルゴリズム、変数増減法による重回帰分析を用いた逐次解析を行い、パス解析によって最適なモデルを目指す。変数増減法は井上・桑山 [6] が作成した R の関数を使用した。

まず、グラフィカルモデリングで求められた偏相関係数行列は図 2，グラフィカルモデリング、SGS アルゴリズムによるグラフはそれぞれ図 3，図 4 のようになった。なお、計算で

必要な基準はそれぞれツールの標準値としたが、グラフィカルモデリングではフルモデルと縮小モデルとの逸脱度に対する p 値は 0.5 であり、偏相関係数は 0.2 である。また、SGS アルゴリズムの独立と判定する偏相関係数は 0.05 とした。

	食費	住居費	年間収入	貯蓄現在高	負債現在高	月消費支出
食費	-1.00	0.00	0.28	0.33	0.37	0.42
住居費	0.00	-1.00	-0.22	-0.42	-0.29	0.57
年間収入	0.28	-0.22	-1.00	-0.31	0.00	0.45
貯蓄現在高	0.33	-0.42	-0.31	-1.00	-0.36	0.32
負債現在高	0.37	-0.29	0.00	-0.36	-1.00	0.00
月消費支出	0.42	0.57	0.45	0.32	0.00	-1.00

図 2: 月消費支出分析の偏相関係数行列

4.1 SGS アルゴリズムの結果考察

図 2 と図 4 に基づいて以下の考察が行える。

- 月消費支出と住居費の偏相関係数値が 0.57 と特に高いことがわかる。
- 月消費支出に対して住居費、年間収入が両方に矢印が付いている。
- 住居費→月消費支出へ向けることは自明である。しかし、月消費支出と年間収入については両方考えなければいけない。なぜなら、偏相関行列から見ても年間収入が大きいほど月消費支出は大きくなるのは理解できないことなく、また月消費支出が大きいほど年間の収入を上げようと努力するという人もいると考えられるからである。
- 貯蓄現在高→年間収入については、貯蓄現在高が高い人は、年間収入も高いというのは関係があるように思われる。
- 貯蓄現在高→住居費では、偏相関係数値が -0.42 と負の相関があることから貯蓄現在高が高い人は住居費が安い傾向にあると考えられる。

両矢印について

SGS アルゴリズムの手順から基本的には両方の矢印が付くことはない。それではなぜ両矢印がつくかという点と数学的アルゴリズムの計算機上での実現の影響と榊原 [3] の R のプログラムの処理の仕様によるものである。図 4 について考察する。

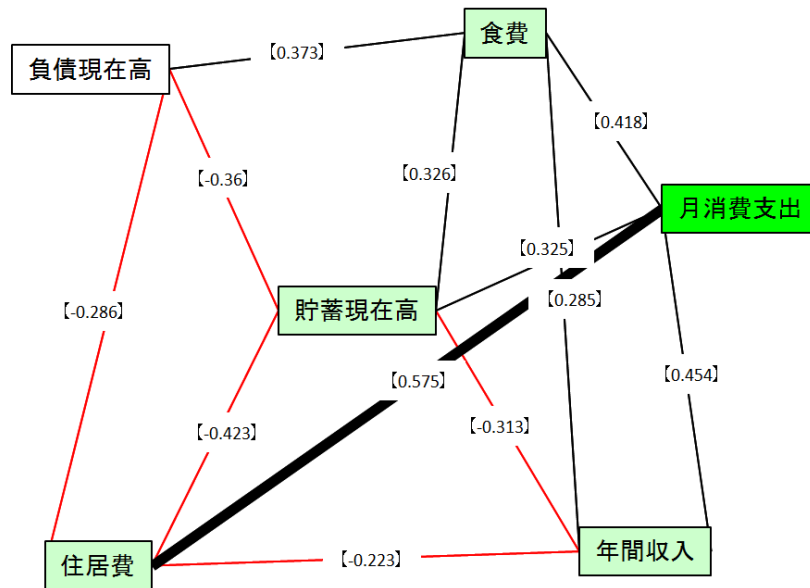


図 3: 月消費支出分析での GM の結果図

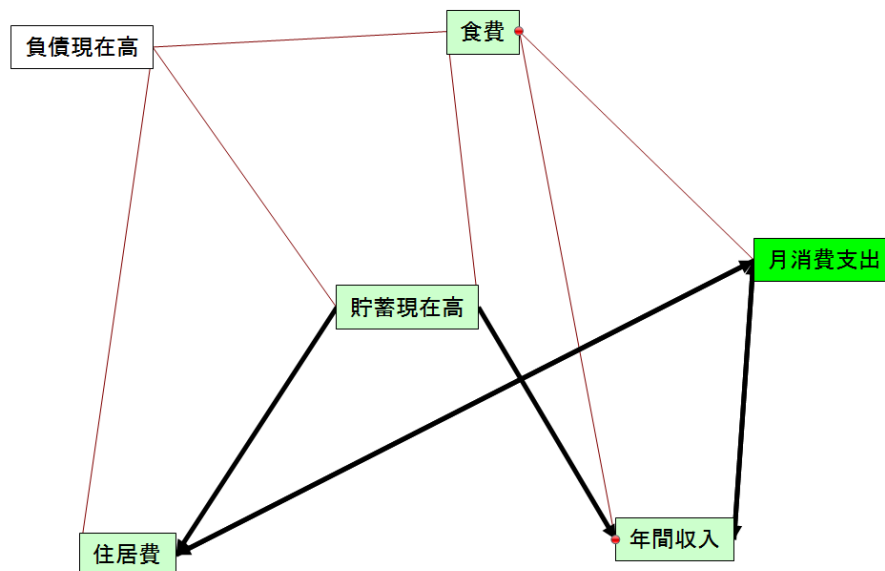


図 4: 月消費支出分析での SGS の結果図

月消費支出と年間収入との線では貯蓄現在高→月消費支出←年間収入と、月消費支出→年間収入←貯蓄現在高の2つを同時に合流と考え、プログラムによって計算しているため両矢印となっている。

同様に、月消費支出と住居費との線では月消費支出→住居費←貯蓄現在高と、住居費→月消費支出←年間収入を同時に計算しているため両矢印の出力結果になっている。

本来、数学的にはこれらの独立の重複は起こりえない。月消費支出と年間収入の場合、貯蓄現在高→月消費支出←年間収入←貯蓄現在高が正しいとすると年間収入を含む変数群で月消費支出と貯蓄現在高が有向分離可能であり、この2つの間に条件付き独立性が成り立つはずだからである。数学的には厳密に独立性が調べられても計算機上ではある一定の基準で判断せざるをえず、その結果このように一見矛盾するグラフを導出することになったと考えられる。また、手順2において無向グラフ K を得るところでも判定基準を用いており、このグラフが間違っている可能性も否定できない。さらに、打ち消しあい問題と呼ばれる本来あるはずの線が見かけ上消えてしまう問題が起こりうることも榊原 [3] で確認されており、その影響もありうる。

一方、榊原 [3] の R のプログラムの処理では手順3をすべて並行して処理するため矢線となった情報を他のチェックでは用いない。万が一、順次矢線にしたものを次は使わないとする仕様に変更した場合、このような事態は避けられるが、どちらの合流が有効とされるかは探索した順序に依存したものとなる。それよりは現在の仕様のままで両矢印が生じた場合は解析者が吟味する方がよりよい結果を導くために適切だと考える。

4.2 変数増減法による重回帰分析を用いた逐次解析

前節の状況を踏まえて変数増減法による重回帰分析を用いた逐次解析を行う。

解析は次のようになった。

- 月消費支出から考えて食費、住居費、年間収入、貯蓄現在高が上流にあると考えられ、図4によって矢印がある年間収入を次に選んだ。
- 年間収入から食費、貯蓄現在高と選択され、図3、図4から食費を次に選んだ。
- 食費から貯蓄現在高、負債現在高、住居費となったが、住居費は図3、図4から関係線がなく、負債現在高を選び、図3から次に、貯蓄現在高という流れで選択をした。

上記の考察から、月消費支出、年間収入、食費、負債現在高、貯蓄現在高という順に逐次解析を行った結果で得られた非巡回的有向独立グラフに榊原 [3] と浅井 [4] によるパス解析を適用すると $AIC: -5.704$, $AGFI: 0.985$, $GFI: 0.997$ となった。修正済み決定係数とも言える $AGFI$ 値が1に近く良いモデルであるが、ここから、貯蓄現在高→負債現在高の矢印を逆向きの負債現在高→貯蓄現在高にすると、 $AIC: -7.057$ と値がさらに小さくなり、 $AGFI: 0.964$, $GFI: 0.993$ となった。その結果が図5である。他の順序に基づく逐次解析もいくつか試したが、これよりよい結果になるものは見当たらなかった。

4.3 パス解析結果の解釈

図5に関して次のようなことが分かった。

- すべての変数の上流は住居費となった。住居にいくらかけるかによって家計の経済が決まってくるということになる。
- 食費から年間収入という関係が一番強く正の係数であった。この向きは不自然なようにも映るが、逆向きなど他の有向グラフを検討しても AIC 値が良くなることはなかった。結果として、収入に応じて食べるのではなく、食べるために働くという構図になる。
- 住居費から負債現在高、負債現在高から貯蓄現在高、貯蓄現在高から年間収入の各線が負の係数となった。住居に金をかけると負債が増え、負債が増えると貯蓄が減るのは自明である。貯蓄が増えると年収が減るのは働かなくても暮らせる層があるせいであろう。

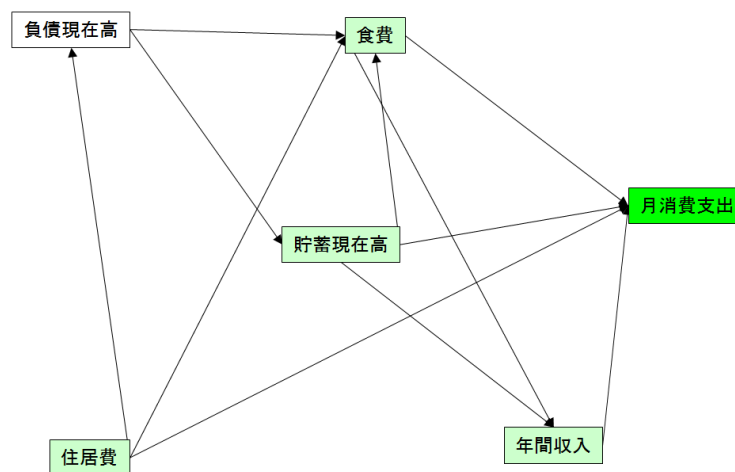


図 5: 月消費支出分析でのパス解析の最終結果図

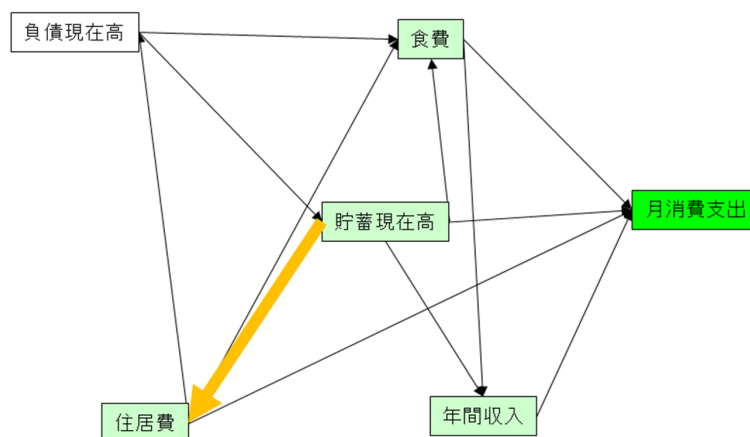


図 6: 月消費支出分析での有向線追加モデル

4.4 SGSの有向線について

貯蓄現在高→住居費の有向線が、SGS アルゴリズムによって取り除かれている結果になったが、図 6 のように線を加えてみたところ、 $AIC: -5.683$, $AGFI: 0.984$, $GFI: 0.997$ と AIC 値が増える結果となった。これによって SGS アルゴリズムについての線が必ずしも良いわけではないということがわかった。

5 考察

最良なモデルを推定する一つの手段として、SGS アルゴリズムを採用することについていくつかの問題点がある。

1 つ目は 4 節で述べた両矢印の問題である。宮川 [2] は、SGS アルゴリズムのルールを適用する変数の順番をランダムに考えるとしており、それによって矢印の結果が変わることを問題点としていた。これを榊原 [3] がプログラムによってランダムではなく一度に計算を行う仕様にした結果、両矢印の結果を出力することになった。このことによって使用者が線の方向を自ら決めることが増えたが、宮川 [2] の結果が変数の順序に依存する問題についてはこの仕様によって解決できていると考えた。しかし、実際データ数が豊富であれば両矢印が出力されることは減るであろうが、今回の例のようにデータが少ない場合は解釈が困難な場合が生じるかもしれないので注意を要する。

2 つ目は、重回帰分析やパス解析につなげる際に SGS アルゴリズムで出力された結果が明らかに原因と結果が逆であったり、SGS での線を増やすとかえってモデルが悪くなってしまう線があるところである。これは、条件付き独立で関係を与えている SGS アルゴリズムの大きな問題点であるといえる。数学的な独立関係を計算機上で実現するにあたってはどうしてもその判断基準の問題から逃れられない。榊原 [3] もこの点について触れているが、変数の数やデータの量に依存するため一律な基準を設けにくい。そのため、すべての関係線が良いモデルになる線として使われることは難しいため、変数選択やその他の統計的手法を使う等をして、SGS アルゴリズムを最良なモデルを目指すための補助を行う道具として使用することが、良い使い方ではないかと考える。

6 おわりに

本研究では、グラフィカルモデリングと SGS アルゴリズムの解析支援ソフトウェアを EXCEL で実行をできるようにし、SGS アルゴリズムとグラフィカルモデリングを融合して因果分析を行うことが目的であった。グラフィカルモデリングと SGS アルゴリズムの支援ソフトウェアの実装はできたが、2 つの連携の部分等、さらに使用者が使いやすくなるような改善点が多くあると考えている。

SGS アルゴリズムの使用については、有向線が実際問題として逆の向きに矢印を向けてしまっていることもあるが、モデルとして良くなる傾向にある線も多く存在するため、SGS アルゴリズムを使用することは良いモデルを出す 1 つの方法論であるといえる。

参考文献

- [1] 宮川雅巳 (1981) : 『グラフィカルモデリング』, 朝倉書店.
- [2] 宮川雅巳 (2004) : 『統計的因果推論』, 朝倉書店.
- [3] 榊原浩晃 (2007) : 『グラフィカルモデリングによる因果推定の研究』, 南山大学大学院数理情報研究科修士論文.
- [4] 浅井悟史 (2012) : 『従業員満足の因果分析に関する研究』, 南山大学紀要『アカデミア』情報理工学編, **12**, 45-55.
- [5] 谷口純一 (2013) : 『グラフィカルモデリングの解析支援ソフトに関する研究』, 南山大学大学院数理情報研究科修士論文, 2013.
- [6] 井上勤・桑山知裕 (2001) : 『S-plus における回帰分析の変数選択関数の作成』, 南山大学経営学部情報管理学科卒業論文.