SGS アルゴリズムとグラフィカルモデリングの支援ソフトに関する研究

M2013SS010 野口良輔
指導教員：松田眞一

1　はじめに

SGS アルゴリズムという確率的にある程度有向グラフを導き出すことのできる手法がある．その手法を統計解析ソフト R 上での関数として実行できるものを榊原 [3] が作成している．グラフィカルモデリングとは, 多変量データの因果関係を分析する方法として開発された多変量解析法の 1 つである．パス解析とは違い, 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を行うことによってその妥当性を検証する事ができ, その統計モデルをグラフによって視覚的に表現することで解りやすく構造を知ることができるものである．本研究でのソフトウェアで解析をする際に Excel ではなく, R での榊原 [3] の関数を用いて計算を行い, Excel によってグラフの出力を行う．谷口 [4] のソフトウェアでは, 無向独立グラフの作成までであったが, 本研究では, SGS アルゴリズムのプログラムを使用することによって, 有向グラフの表示を行い, 例題のデータを用いて解析を行い, 比較検討を進めていく．

2　実装に用いたプログラミング言語

2.1　R 言語

R 言語とは, オープンソースのフリーソフトウェアで, 統計解析向けプログラミング言語とその開発実行環境である．統計解析の分野では様々な分析方法が存在しており, それらの多くは複雑な数式から得られ, プログラミング言語で実装するには大量のコードを記述する必要がある．しかし, R 言語ではそれらの分析処理を 1 つの関数として解析を行うことができるため, 非常に便利な言語である．本研究では, 開発支援ソフトで扱われる手法である SGS アルゴリズムとグラフィカルモデリングについて榊原 [3] が作成した R 言語のプログラムによって計算を行っている．

2.2　VBA

VBA(Visual Basic for Applications) とはマイクロソフト社製の Microsoft Office シリーズに搭載されているプログラミング言語である．VBA を使用することで, 定型業務を自動化することができ, 様々なプラグインを組み込むことで機能をカスタマイズできる．本研究では, 一般的によく用いられているであろう Microsoft Office シリーズの Excel を用いることで, この支援ソフトを利用するユーザに簡易に使ってもらえるよう VBA での支援ソフトの作成を行った．

3　グラフィカルモデリング

グラフィカルモデリングは一本ずつの線の切断（以後線断とよぶ）を相関係数から計算される偏相関係数の値を見ることで条件付き独立かどうかを判断し, グラフとして線を結ぶことでモデルの推定を行う方法である．その際に無向グラフのフルモデルを初期状態として, 初期状態とグラフの線断を行ったグラフと比較することでグラフ構造が大きく変化していないかを確認しながら進める．(宮川 [6], 榊原 [3] 参照)

3.1　共分散選択と評価

モデルの線断を行う際に使われるのが共分散選択である．
標本相関係数を $R = r_{ij}$ として逆行列を $R^{-1} = r^{ij}$ としたとき, 以下の偏相関係数が求められる．

$$R_{ij \cdot rest} = -\frac{r^{ij}}{\sqrt{r^{ii}}\sqrt{r^{jj}}}$$

それらから 0 に近い数値を選択して小さいものから順番に 0 として条件付き独立関係を与えていく共分散選択を行い, それから Dempster の定理を用いて縮小モデルを作成し, 多変量正規分布を仮定した尤度関数から逸脱度を求めて検定を行う．(豊田 [5] 参照)

グラフィカルモデリングの手順は以下に示す．(宮川 [6] 参照)

グラフィカルモデリングの手順

1. 偏相関係数値の一番小さい値をとる変数の番号の組を *(i, j)* とする．
2. 与えられた (i, j) を使い $(a, b) \neq (i, j)$ としたとき $\sigma_{ab} = s_{ab}$, $\sigma^{ij} = 0$ のように条件付き独立とする．
3. 条件付き独立となったすべての偏相関係数値が *0* となった場合は *5* へ進む
4. 手順 *2* を行うことによって他に条件付き独立とした値が極端におおきくなれば, 大きくなった項目を独立した手順まで戻る．増加した量が少しの場合はその項目 *(i, j)* を元に手順 *2* をもう一度行う．
5. フルモデルと縮小モデルの評価をしてモデルが大きく外れていないかを判断する．もし評価が悪いようであれば条件付き独立にするのは間違っていたと判断し, 切らないことにする．
6. 条件付き独立とする候補が未だあるようであれば *1* に戻る, 無いようであれば終わる．

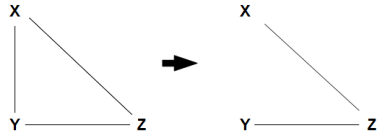

図 1　手順 2

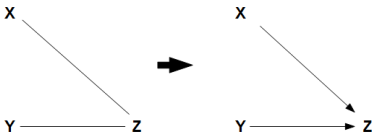

図 2　手順 3

4　SGS アルゴリズム

SGS(Sprites, Glymour, Scheies) アルゴリズムとは, 巡回していないモデルのみを扱う数学的アルゴリズムであり, 条件付き独立によって導き出す手法である. グラフィカルモデリングでは無向グラフが出力されるが, SGS アルゴリズムでは有向線を含む混合グラフが出力される. (宮川 [8], 榊原 [3] 参照)

SGS アルゴリズムの手順は以下に示す. (宮川 [8] 参照)

SGS アルゴリズムの手順

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, $V\backslash(X_i, X_j)$ のある部分集合 S(空集合の時もある) を与えた時に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. (図 1 参照)
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^* を与えた時, 条件付き独立になるような S^* 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 (図 2 参照).
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, $X_k \rightarrow X_j$ と矢印をつける. (図 3 参照)
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 に有効道があり, 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. (図 4 参照)
6. この 4, 5 の手順を矢線をつける辺がなくなるまで続ける.

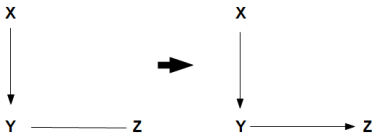

図 3　手順 4

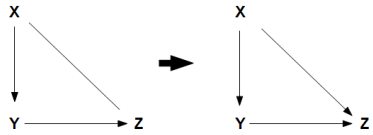

図 4　手順 5

4.1　問題点

SGS アルゴリズムにおける線断問題がある. 真のデータでは条件付き独立となる場合の値は 0 であるが, しかし実際のデータを扱う場合 0 より若干大きくなってしまう. (榊原 [3] 参照) よって解析を行う際は解析者の判断で打ち切りの基準を定めなければいけない. ここで榊原 [3] が行った線断基準を定めるシミュレーションの結果である, SGS アルゴリズムの打ち切り基準表から妥当だと思われる 0.05 という数値を偏相関係数値の打ち切り基準とし本研究での支援ソフトで使用する.

5　支援ソフトについて

浅井 [1] の研究によって作成されたパス解析を支援するソフトウェアと榊原 [3] によって作成されたグラフィカルモデリング, SGS アルゴリズムの R 関数を組み合わせてグラフィカルモデリングと SGS アルゴリズムの結果を別 Book で出力できる支援ソフトの作成を行う. なお, グラフィカルモデリングのプログラムは谷口 [4] より浅井 [1] をベースにして抜本的改修を行った. 本研究で試みる支援ソフトの実行プロセスを以下に記載する.

ソフトウェアの実行プロセス

1. Excel で R の実行パスをユーザに指定させる
2. データをテキストファイルに入力後に Excel でデータのファイルのパスを保存する.
3. R の実行命令文を作成し, バッチコマンドとして R を実行し, GM, SGS アルゴリズムの計算を行う.
4. R の計算結果を R で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出力させる.
8. Excel でグラフを作成する.

6　支援ソフトの使い方

6.1　初期設定

GM, SGS フォームから, ボタン [R 初期設定] を選択し, ウインドウが現れる. R のディレクトリ内の R.cmd を選択し R を Excel(VBA) によって動せるようにする. R.

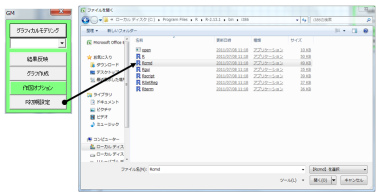

図 5　R のディレクトリの指定

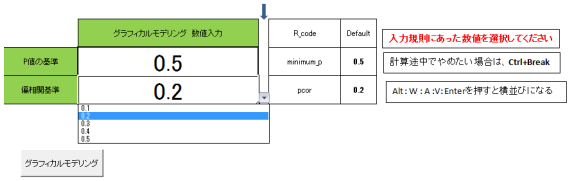

図 6　GM 入力

cmd のディレクトリ指定をするときには図 5 を参照されたい.
※初期設定は最初の 1 回のみでよい.

6.2　R の実行コード入力について

Sheet[Input] から GM, SGS 支援ツールそれぞれ GM：P 値の基準, 偏相関基準, SGS：線断基準, 全通りするか否かを入力することで R の命令文を変更できる. 入力画面については図 6 を参照されたい. SGS の入力についても同様になっている.

6.3　テキストの選択

▼のボタンを押すことで, 支援ツールフォルダの● table 内のテキストファイルが一覧されるようになっている. このデータ選択の処理は, 浅井 [1] の研究成果であるパス解析支援ツールがヒントになっている.
データ選択の画面については, 図 7 を参照されたい. その後, ボタン [グラフィカルモデリング] または, ボタン [SGS アルゴリズム] を選択し R の実行結果をテキストへ出力させるところまで行う. ※分析対象のデータはテキストファイルを指定する, またデータのそれぞれの変数の名前は必ずつけるようにし, 最後の列を目的変数とする事を仕様としている.

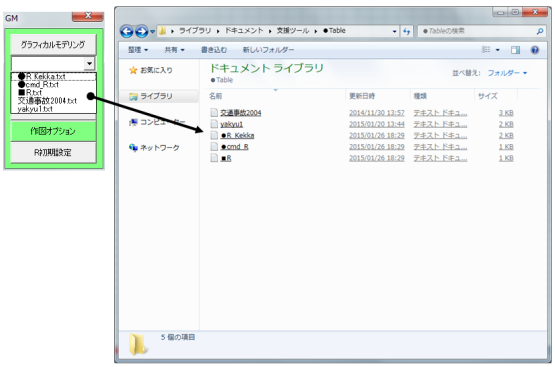

図 7　テキストの選択

6.4　結果反映

R の解析後, Sheet へ反映させるために各フォームの [結果反映] を選択することで, 変数の数のカウント→それらをスペース区切り→変数の名前の読み取り→ Sheet へ反映する. また, GM 支援ツールで出力された偏相関係数行列は同ツール内の Sheet[Matrix] に出力される.

6.5　作図

榊原 [3] の R の関数は線や因果の向きを 0 – 1 変数の行列の形式で与えるので, それに基づいて Excel 上で作図を行う. 作図の処理は浅井 [1] の研究成果であるパス解析支援ツールがヒントになっている.

6.6　作図したグラフの同期

2 つの Book から解析を行っているため, それぞれのグラフを同じような形にする必要がある.
作業プロセス：[同ディレクトリに支援ツールの有無を確認] → [別シートが計算後であるかを確認] → [座標の同期] → [グラフの出力] という形で動かしている.

7　データ解析例

2004 年都道府県別消費支出についてのデータを用いて目的変数を [月消費支出] とおいて, [食費], [住居費], [年間収入], [貯蓄現在高] 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, 変数増減法を用いた重回帰分析による逐次解析を行い, 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 [2] が作成した zougenhou という R の関数を使用する.
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 9, 図 10 のようになった.

7.1　SGS の考察

SGS アルゴリズムによって出力されたグラフの矢印の付いた線は図 10 になった.
・[月消費支出] と [年間収入], [住居費] がそれぞれ両矢印の線になっている.
・[年間収入] → [月消費支出] では, 年間収入が高いほど月消費が高いという意味であり, [月消費支出] → [年間収入] では, 月消費が高いからこそ年間収入を増やすという意味も考えられるため, 両方の矢印のどちらかを採用する余地があるといえる.
・[月消費支出] と [住居費] では, [月消費支出] の中に [住居費] が包括されている関係であるので, [住居費] → [月消費支出] の線を採用する.
・[貯蓄現在高] → [年間収入], [貯蓄現在高] → [住居費] の 2 つの線は偏相関係数値が負であるため, それぞれの線は負の相関関係という結果になった. 出力された偏相関係数行列については, 図 8 を参照されたい.

7.2　パス解析結果

これらを念頭に置いて逐次解析からパス解析を行った結果, AIC : −7.057, $AGFI = 0.985$, $GFI = 0.997$ と

	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 8　偏相関係数値

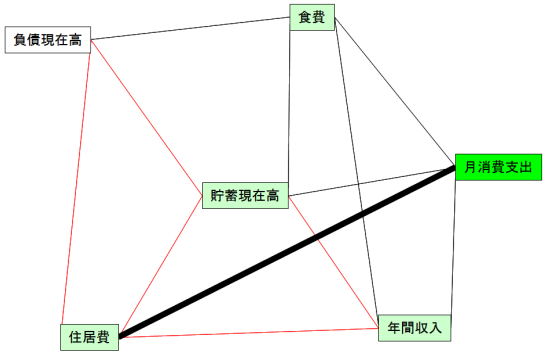

図 9　GM の結果図

なった. 結果については図 11 を参照されたい.
また, SGS によって出力された [貯蓄現在高] → [住居費] を追加した結果 (図 12 参照)
AIC : −5.683, $AGFI = 0.984$, $GFI = 0.997$ とパス解析での修正済み決定係数における AGFI 値が下がり, AIC 値が正に大きくなってしまったので, SGS によって出力された線をすべて参考することは難しいと考えた.

8　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を作成し, 2 つのグラフの結果から解析をより良いものにするヒントになるかを実際のデータを用いて理解する目的であった.
2 つの支援ソフトの実装はできたが, 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

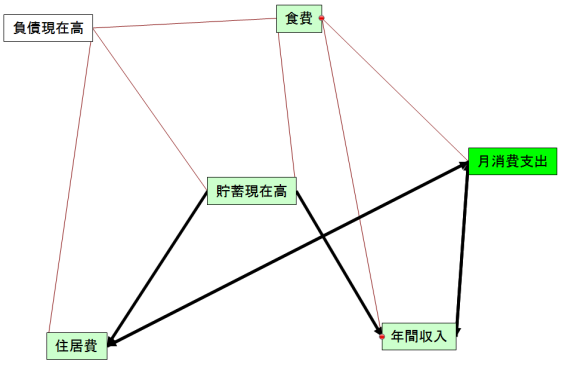

図 10　SGS の結果図

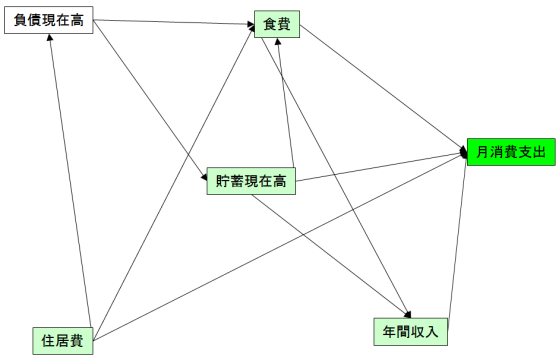

図 11　パス解析の最終結果図

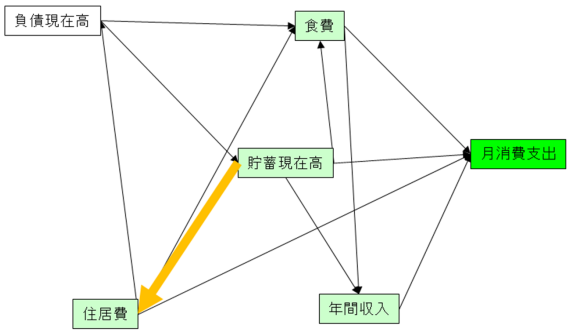

図 12　追加したモデル

と考えている.
SGS アルゴリズムを導入することによって, 有向線が逆の向きに矢印を向けてしまっていることもあるが, モデルとしてよ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探す 1 つの方法論であるといえる.

参考文献

[1] 浅井悟史 :『従業員満足の因果分析に関する研究』. 南山大学大学院数理情報研究科修士論文, 2013.

[2] 井上勤・桑山知裕 :『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, 2001.

[3] 榊原浩晃 :『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, 2007.

[4] 谷口純一 :『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.

[5] 豊田秀樹 :『共分散構造分析　入門編-構造方程式モデリング』. 朝倉出版, 1998.

[6] 宮川雅巳 :『グラフィカルモデリング』. 朝倉書店, 東京, 1981.

[7] 宮川雅巳 :『グラフィカルモデリングの実際』. 朝倉書店, 1999.

[8] 宮川雅巳 :『統計的因果推論』. 朝倉書店, 2004