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

M2011MM071 谷口純一
指導教員：松田眞一

1　はじめに

本研究は，Excel 用プログラミング言語である VBA を用いて，統計解析の手法であるグラフィカルモデリングの無向独立グラフの出力までを，Excel で実行するシステムの実装を目的とする。

本研究は，先行研究である浅井 [1]，榊原 [3] の研究を基にしている。グラフィカルモデリングは，榊原 [3] の研究で統計解析ソフト R を用いた解析が提示されていた。R は市販されているソフトウェアでなく，フリーウェアであるので，個人でインストールして利用する事ができる。本研究の以前まで，榊原 [3] の関数を利用することにより，R をある程度使いこなせる利用者のみグラフィカルモデリングの分析を行うことができた。

グラフィカルモデリングは，変数相互の関係をグラフに図示する事により，視覚的に変数間の関係を把握することができる。しかしながら，榊原 [3] の研究では，グラフィカルモデリングの計算結果のみを出力し，変数間のグラフの図示は行わない。一方，浅井 [1] においてパス解析を Excel 上で行う研究がなされており，その際，自動的にパス図の作成がなされていた。

本研究は，先行研究をさらに発展させ，Excel でデータを読み取り，R での計算を経由して，Excel でグラフィカルモデリングの無向独立グラフの出力までを行う。これは，R にある程度詳しくない初学者でも，データの読み込みから変数間のグラフを作成できるという点に研究を行う意味がある。先行研究のプログラムの一部を用いて，表計算ソフト Excel で，R と連動したグラフィカルモデリングの解析支援ソフトの実装を行う。本ソフトウェアは，Windows と Microsoft Office2010 Excel で動作を行う事を考えている。また，統計解析ソフト R はバージョン 2.14.0 を利用している。

2　ソフトウェアの実装に用いた言語について

2.1　R について

R 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。本研究では，グラフィカルモデリング解析支援ソフトの中からグラフィカルモデリングの計算を行うために利用している。統計解析ソフト R はフリーウェアであるので，適宜ダウンロードし，インストールを行うことで，利用する事ができる。

2.2　VBA について

VBA(Visual Basic for Applications) は Microsoft 社で商標登録されている Microsoft Office シリーズに標準で搭載されているプログラミング言語である。VBA は，Excel に標準で付属されている言語で，Excel 内で他のソフトウェアをインストールする事無く，プログラムを作成する事ができる。本研究で作成するソフトウェアも，Excel 内で実行するプログラム言語として VBA を利用している。解析支援ソフトの利用者は，Windows の環境に Excel を導入する事で作成するソフトウェアを利用する事ができる。VBA で作成したプログラムは，Excel ファイルの一部として格納される。

3　グラフィカルモデリング

グラフィカルモデリングは，線の切断 (以下では線断と呼ぶ) を行うことで，以下の手順のようにモデルの推定を行う。

初期状態をフルモデルの無向グラフとし，グラフの線断は，初期状態のモデルと比較し，構造が大きく変化していないかを確認しながら進める。(榊原 [3]，宮川 [5] 参照)

グラフィカルモデリングは，相関係数から計算した偏相関係数の値から，変数間に関係があるかを考えた上で，グラフにより変数間のモデルを表現する。そして，変数間の関係を推定し，グラフの線断を行う。

分析対象のデータの標本相関係数を計算し，求めた相関係数から，偏相関係数を求める。母相関係数を σ_{ij} とし相関行列の逆行列を σ^{ij} とした際に，偏相関係数 $\sigma_{ij\cdot rest}$ は以下のようになる。

$$\sigma_{ij\cdot rest} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}}\sqrt{\sigma^{jj}}}$$

グラフィカルモデリングの手順

1. 偏相関係数値の最も小さい値となる変数の組を (i,j) とする。
2. (i,j) にて，$\sigma_{ij\cdot rest} = 0$ のように条件付き独立とする。
3. 条件付き独立としたすべての偏相関係数値が 0 となるように繰り返し偏相関係数行列を推定する。
4. フルモデルと縮小モデルとの差を評価をして，モデルが大きく外れていないかを判断する。仮に，モデルの適合度が悪い場合は，条件付き独立としていたことが間違っていたので，変数間の線断を行わないで終了する。
5. 条件付き独立とする候補が残っている際は，手順 1 に戻る。候補が残っていなければ終了する。

共分散選択で Dempster の定理より (i,j) のとき，$\sigma_{ij\cdot rest} = 0$ ならば，変数間の線断により縮小モデル (RM:Reduced Model) を導く事ができる。(宮川 [6] 参照)

そのため，各相関係数から偏相関係数行列を求め，偏相関係数が限りなく 0 に近いものを選択する。偏相関係数が限りなく 0 に近い部分は，変数間が独立であること

を示すので，条件付き独立関係を意味する。そして，最終的に導き出された縮小モデルのグラフが無向独立グラフとなる。(榊原 [3]，宮川 [5] 参照)

3.1 線断基準について

グラフィカルモデリングで一般的に作成されたモデルは，フルモデルとの差の適合度検定を行い，P 値により線断の判断を行っている。(宮川 [5] 参照)

しかしながら，P 値の打ち切り基準は，一意に決定しているのではなく，おおよそ 0.5 程度を基準としている。(宮川 [5] 参照)

グラフィカルモデリングは，パス解析と異なり，どの変数間に線が存在し，どのようなモデル図になるかを計算により求める。グラフィカルモデリングで打ち切り基準の P 値をどの程度の値にするかは，モデル図を作成する上で，非常に重要な問題点であり，榊原 [3] は，打ち切り基準の問題として着目し，どのような打ち切り基準が良いかをシミュレーションで示した。それにより，本研究の解析支援ソフトは，P 値の打ち切り基準を 0.03 とした。打ち切り基準をシミュレーションで求める方法については，榊原 [3] を適宜参照されたい。

3.2 合流問題について

グラフィカルモデリングで，因果合流が存在する際に，説明変数間に擬似的な直接効果の関係が生じる。このため，直接的な関係がない場合にも，擬似的に直接効果と同様の線が生じてしまう。また，逆に線があるのに打ち消しあって消えることもある。(榊原 [3] 参照)

4 パス解析とグラフィカルモデリングについて

浅井 [1] の研究は，従業員満足を分析対象としたパス解析による考察を行っている。パス解析は，事前に変数間の関係の方向性を決定し，「0」,「1」でどの方向に線が存在するかを指定している。小島 [2] でも，変数同士の方向性が事前に明らかでない状況が生じうるが，パス解析は，分析対象の変数の相互関係がある程度，事前に想定できる状況が必要であると述べている。しかしながら，実際の分析は，分析対象の関係性が全く明らかでない状況がしばしば生じる。グラフィカルモデリングは，そのための方法論であり，分析対象の変数間で，どの変数間に線が存在し，どのようなモデル図になるかを計算により決定する。本研究は，分析対象を定量的な視点から議論し，モデル図を作成する。

5 ソフトウェアの処理の流れ

実装するソフトウェアは，以下の部分ごとに機能を組み合わせてソフトウェア全体の実装を行う。なお，次に示すソフトウェアの処理のうち，1 と 2 と 3 から 7 と 8 は，それぞれのボタンで操作できるようにした。特に 3 から 7 は自動で行える。

ソフトウェアの処理の流れ

1. Excel で R の実行パスをユーザに指定させて保存
2. データをテキストに入力後，Excel でデータのファイルパスを保存
3. R 実行命令文のテキストファイルを作成し，バッチコマンドとして R を実行し，グラフィカルモデリングの計算を行う
4. R の計算結果を R でテキストファイルに書き出す
5. Excel で計算結果のテキストファイルが作成しているかを確認する
6. Excel で計算結果を入力する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を読み込む (計算結果はセル区切りにする)
8. Excel で無向独立グラフを作成する

6 ソフトウェアの仕様について

6.1 R のパスの保存について

R の実行パスを保存するように実装した。(浅井 [1] 参照)

6.2 データの入力について

データの入力は，テキストにデータを入力後，テキストのファイルパスを登録するように実装した。

6.3 R 実行命令文の作成と呼び出しについて

R でのグラフィカルモデリングの計算命令文を図 1 のように Excel で作成し，R の呼び出し処理を実装した。

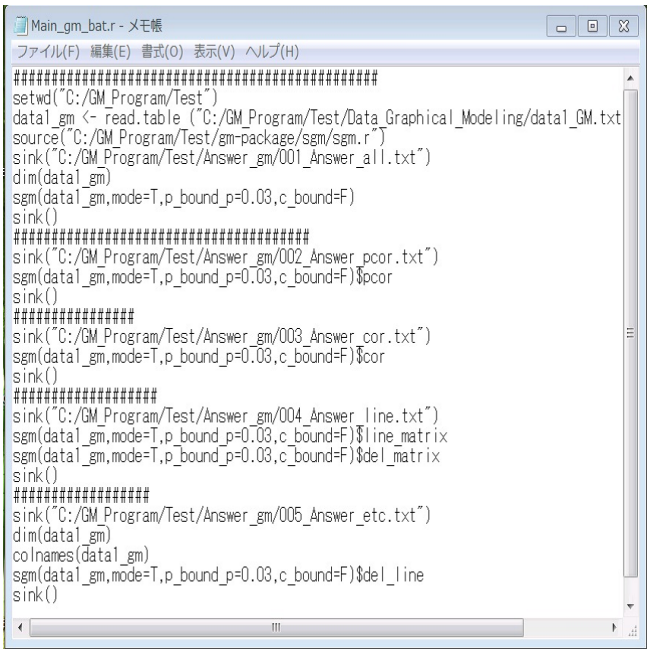

図 1 R の呼び出し処理

6.4　R の計算結果について

R の計算結果をテキストファイルに出力するよう実装した。計算結果の出力は辺行列を含んでいる。辺行列は，分析対象の変数間に線が存在する場合は「1」，変数間に線が存在しない場合は「0」となっている。R の実行は，浅井 [1] の結果を利用した。

6.5　計算結果ファイル作成確認について

計算結果のテキストファイルが作成しているかを確認し，計算結果のファイルが作成されていない場合は，処理エラーとして，その後の処理を止めるように実装した。処理方法は，浅井 [1] を参考にした。計算結果のファイルが作成されている場合は，データ読込用に，Excel のワークシートを初期化する。

6.6　Excel で計算結果の読み取りについて

Excel でデータの計算結果の読み込みを図 2 のように実装した。

	A	B	C	D	E	F	G
2	$cor						
3		圧力	温度	酸度	粘度	収率	
4	圧力	1	0.44159	0.3652913	−0.580444	−0.5385145	
5	温度	0.44159	1	0.5511066	0.1112874	0.1677515	
6	酸度	0.3652913	0.5511066	1	0.455574	0.4857036	
7	粘度	−0.580444	0.1112874	0.455574	1	0.9561449	
8	収率	−0.5385145	0.1677515	0.4857036	0.9561449	1	
9							
10	$pcor						
11		圧力	温度	酸度	粘度	収率	
12	圧力	−1	3.64E−01	8.14E−01	−4.19E−01	−0.2591002	
13	温度	0.364059	−1.00E+00	2.00E−16	−3.96E−06	0.2077577	
14	酸度	0.8138211	2.59E−16	−1.00E+00	3.78E−01	0.2616692	
15	粘度	−0.4192926	−3.96E−06	3.78E−01	−1.00E+00	0.719302	
16	収率	−0.2591002	2.08E−01	2.62E−01	7.19E−01	−1	
17							
18	$del_line						
19		[1]	[2]				
20	[1,]	2	4				
21	[2,]	2	3				
22							
23	$del_matrix						
24		[1]	[2]	[3]	[4]	[5]	
25	[1,]	0	0	0	0	0	
26	[2,]	0	0	1	1	0	
27	[3,]	0	1	0	0	0	
28	[4,]	0	1	0	0	0	
29	[5,]	0	0	0	0	0	
30							
31	$line_matrix						
32		[1]	[2]	[3]	[4]	[5]	
33	[1,]	0	1	1	1	1	
34	[2,]	1	0	0	0	1	
35	[3,]	1	0	0	1	1	
36	[4,]	1	0	1	0	1	
37	[5,]	1	1	1	1	0	
38							

図 2　Excel でのデータ読み取りについて

6.7　グラフ出力について

Excel で読み込んだ計算結果から，自動で無向独立グラフを出力するように実装した。この処理も浅井 [1] を参照した。

7　解析例について

7.1　化学工程データの例

化学工程の収率と製造条件に関するデータの分析を行う。(圓川 [4] 参照)

データの変数項目は圧力，温度，酸度，粘度，収率で，目的変数を収率として，収率以外の変数項目が，収率にどのような影響を与えるかを考える。

圓川 [4] は，説明変数間に相互関連があり，多重共線性が起こった因果モデルを説明として与えた。変数間の相互関係がどの程度あるかを考える。

	圧力	温度	酸度	粘度	収率
圧力	1	0.44	0.36	−0.58	−0.53
温度	0.44	1	0.55	0.11	0.16
酸度	0.36	0.55	1	0.45	0.48
粘度	−0.58	0.11	0.45	1	0.95
収率	−0.53	0.16	0.48	0.95	1

表 1　化学工程データの相関係数行列

表 1 の変数間の相関係数行列より，収率と粘度は相関関係が非常に強い。また，表 2 で，圧力と酸度，粘度と収率の偏相関係数の値が大きい。しかしながら，表 2 のみで，変数間の構造を考える事は難しい。

	圧力	温度	酸度	粘度	収率
圧力	—	0.43	0.83	−0.34	−0.33
温度	0.43	—	−0.12	−0.12	0.39
酸度	0.83	−0.12	—	0.35	0.29
粘度	−0.34	−0.12	0.35	—	0.72
収率	−0.33	0.39	0.29	0.72	—

表 2　化学工程データの偏相関係数行列

解析支援ソフトにて作成した無向独立グラフは図 3 のようになった。

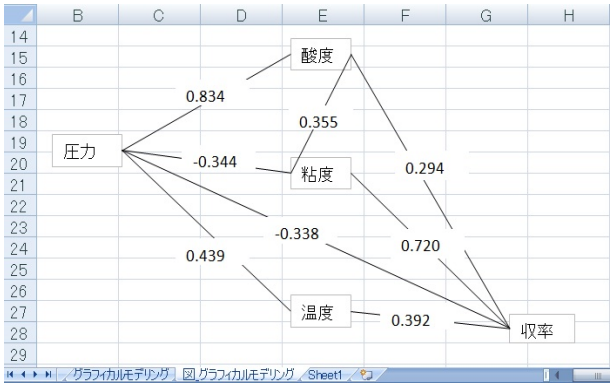

図 3　化学工程データのグラフ

グラフィカルモデリングのグラフ結果より，収率以外の変数項目である圧力，酸度，粘度，温度は，どの変数項目も収率に影響を与えている。その中で，収率に一番影響を与えているのは，粘度となる。また，圧力，粘度の組と圧力，収率の組で，偏相関係数が負の符号であるので，圧力が増えると粘度と収率が下がる。そして，偏相関係数が非常に高い組である粘度と収率の間に，強い相関関係がある。グラフィカルモデリングの後に，因果の向きを推定し，浅井 [1] のパス解析支援ツールで，パス解析を行ったところモデルのあてはまりの良さを示す AIC 値は −1.977，決定係数に相当する AGFI 値は 0.441 であった。AIC 値は，より小さい値が良く，AGFI 値は，より大きい値が良い。ここで，より良いパス解析のモデル図の探

索を行った。グラフィカルモデリングのモデルを基準として，線の線断のみでより良いパス図を探索した結果は，AIC 値は −6.897，AGFI 値は，0.704 となった。これは，圧力と収率，温度と収率，酸度と収率，圧力と酸度の変数間の線の線断を行い，図 4 のようになった。

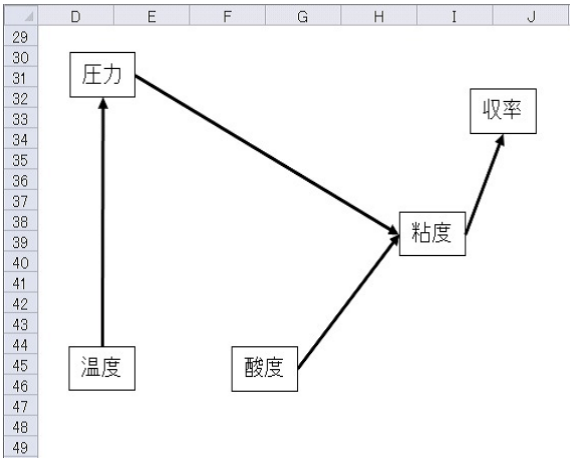

図 4　化学工程データのパス図

宮川 [5] は，圓川 [4] から，引用した同じ解析データについて，パス図を結果として示した。宮川 [5] のパス図の AIC 値は −6.897，AGFI 値は 0.704 であったが，パス図は異なった。

7.2　陸上競技データの例

1995 年日本陸上競技に関するデータの分析を行う。(宮川 [5] 参照)

データの変数項目は，100m，走幅跳び，400m，110mH，1500m といった走る速さに関する 5 種目の陸上競技種目である。データのサンプルサイズである陸上競技を測定した人数は，50 人である。

表 3，表 4 より値に差がなく，それぞれの変数でどのような関係かを考えることは難しい。

	100m	走り幅	400m	110mH	1500m
100m	1	−0.22	0.46	0.35	−0.17
走り幅	−0.22	1	−0.16	−0.24	0.01
400m	0.46	−0.16	1	0.46	0.20
110mH	0.35	−0.24	0.46	1	0.04
1500m	−0.17	0.01	0.20	0.04	1

表 3　陸上競技データの相関係数行列

	100m	走り幅	400m	110mH	1500m
100m	—	−0.18	0.43	0.11	−0.32
走り幅	−0.18	—	0.08	−0.19	−0.10
400m	0.43	0.08	—	0.35	0.32
110mH	0.11	−0.19	0.35	—	−0.03
1500m	−0.32	−0.10	0.32	−0.03	—

表 4　陸上競技データの偏相関係数行列

解析支援ソフトで出力された無向独立グラフを図 5 に示す。グラフの出力の結果から，仮に，400m を基準とした場合に 100m，1500m，110mH，400m 走のタイムが遅い人は，100m，1500m，110mH のタイムが遅くなる。100m と 1500m は，偏相関係数が負の符号となっているが，これは，短距離走と長距離走の違いではないかと考えれる。5 種目の中で，走り幅跳びに注意してみると 400m や 1500m とは線が結ばれていない。走り幅跳びは，長い距離を走る 400m や 1500m といった種目でなく，100m や 110mH といった短距離走の影響がよりあるからではないかと考えられる。宮川 [5] でも，走るのみの種目と走り幅跳びでは，異なる分類となっている。また，100m と 1500m の間には，スピード型と持久力型に区別されると指摘されている。

宮川 [5] にて，示されるグラフィカルモデリングの図では，走り幅跳と 400m の変数間に線が存在する点が，図 5 と異なっている。

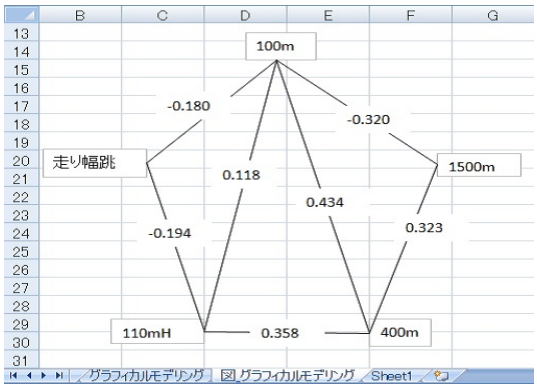

図 5　陸上競技データのグラフ

8　おわりに

本研究を通して，表計算ソフト Excel と R の連動したソフトウェアの処理の流れをもとに，Excel でグラフィカルモデリングを実行できる環境を VBA を用いて実装できた。また，実際に解析を行う上で，ソフトウェアの使い勝手や解析者が解析を行いやすいソフトウェアかを考える事もできた。

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，2012.

[2] 小島隆矢:「Excel で学ぶ共分散構造分析とグラフィカルモデリング」，オーム社，2003.

[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] 圓川 隆夫:「多変量のデータ解析」，朝倉書店，1988.

[5] 宮川雅巳:「グラフィカルモデリング」，朝倉書店，1997.

[6] 宮川雅巳:「統計的因果推論」，朝倉書店，2004.