R によるパス解析の実現とその応用

2001MM073 榊原 浩晃
指導教員　松田 眞一

1　はじめに

現在、アンケートなどの多変量データを解析するにあたって、重回帰分析のようにそれぞれの目的変数への影響度を知るだけでは不十分なときがある。つまり直接影響を及ぼすだけでなく、間接的に説明変数を通り目的変数に影響を及ぼしているかもしれないということを考える必要があるということだ。直接の効果は正の値になっていても総合的に見ると負になる可能性もある。本卒業研究ではこのような総合的な影響力を見つけ出す方法を学び、統計解析ソフト R を用いて解析できる用にプログラムを作成する。

2　パス解析

2.1　パス解析とは

因果の向きを矢線を使って表現し、直接関係のある変数同士を結ぶ。そのときに矢線上に影響力の強さの数値を書く。最終的に完成したモデルをパス図と呼び、変数同士の関連を考慮しながら解析を行う。パス図の良い所は重回帰分析のや数量化 I 類など説明変数間に関係があると問題が生じている場合、つまり多重共線性の問題の場合でも積極的に解析が行えることである。A と B は

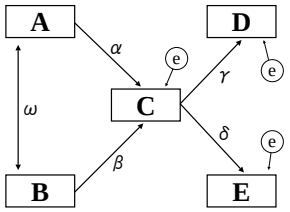

図 1　サンプルモデル

外生変数と呼び、外生変数間は相関を考えるということで両側矢印で結ぶ。A と C においては直接関係があるので直接効果と呼ぶ。B・C・E においては B から E への間接効果と呼び、B と E の相関があるが直接影響を及ぼしているのではなく、C を通って影響を及ぼしている。A・B・C においては、C への因果合流と呼び、A と B の相関が無いときも C が固定された場合 A と B の間に相関関係が生まれる性質がある。C・D・E 間においては D と E にそれぞれ相関があり、C が原因となり影響を及ぼしているので結果的に D・E 間に相関が生まれる。この相関を疑似相関と呼ぶ。このとき C の値を固定した場合 D と E は条件付独立となる。

また、疑似相関や間接効果の相関はパス係数の積にて求まる。最後に総合効果というものがあり、直接効果と間接効果のルートのそれぞれの値の和にて求まる。この総合効果は説明変数が目的変数に与える本当の影響力を知ることができる。

2.2　母数の推定方法

本研究では母数の推定方法は最尤推定法を用いて推定をする。最尤推定法は標本データが多変量正規分布に従っていると仮定し、自由母数（推定すべきパス係数、残差・外生変数の分散共分散）を未知数として、求められる P 値が最大となるときの母数を最尤推定値として採用する手法である。

つまり、多変量正規分布に従っていると仮定したので、確率密度関数は

$$f(\boldsymbol{x}|\boldsymbol{\mu\theta}) = (2\pi)^{\frac{-n}{2}}|\boldsymbol{\Sigma(\theta)}|^{\frac{-1}{2}}\exp[\frac{-1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma(\theta)}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \tag{1}$$

となる。ただし $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$、$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)'$、共分散行列を $\boldsymbol{\Sigma(\theta)}$ とする。ここで、f の対数をとり母数に関係の無い定数項を加えることによって最尤推定量のための目的関数 f_{ML} の値は

$$f_{ML} = tr(\boldsymbol{\Sigma(\theta)^{-1}S}) - \log|\boldsymbol{\Sigma(\theta)}^{-1}\boldsymbol{S}| - n \tag{2}$$

となる。よってこの値が最小値を取る θ を最尤推定値として採用する。

2.3　モデルの推定

モデルを推定する一つの方法として、適合度検定を行う。観測データが多変量正規分布であることを仮定し、モデルが正しければ式 (2) より

$$\chi^2 = (N-1)f_{ML} \tag{3}$$

が自由度 $df = \frac{1}{2}N(N+1) - p$ (p は θ の次数) の χ^2 分布に近似的に従うことを利用し、適合度検定を用いてモデルの推定を行う。

また、回帰分析における決定係数と同じ意味をもった GFI や自由度の制約した AGFI を参考にするのも有効である。

$$GFI = 1 - \frac{tr\{(\Sigma^{-1}(S-\Sigma))(\Sigma^{-1}(S-\Sigma))'\}}{tr\{(\Sigma^{-1}S)(\Sigma^{-1}S)'\}}$$

$$AGFI = 1 - \frac{N(N+1)(1-GFI)}{2df}$$

3　モデリング

3.1　グラフィカルモデリング (GM)

グラフィカルモデリングとは多変量データの構造をより分かりやすくするものである。つまり、それぞれ影響を及ぼしあっている変数を線で結び、変数間の関係が独立であれば線を切るといった操作を行い多変量データを

モデリングし、因果の向きは考えないモデルである無向独立グラフを作成することである。つまり間接効果や疑似相関の関係を見つけだし、不要となる線を消していくという作業を行うことである。また、モデリング後の図を使いパス解析を行うことにより変数間の影響の度合いを数値化し現状の把握や問題解決に役立たせることができる。ここでグラフィカルモデリングを行う手順は

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手順 1 を繰り返す。

この作業を行いグラフィカルモデリングを行う。図 1 に対応する無向独立グラフは図 2 である。

3.2 グラフィカル連鎖モデリング

変数が多い状態でグラフィカルモデリングを行うと、因果の向きを推定し辛くなる。つまり、矢線なのか因果合流のための線なのかを決めることが難しくなってくる。そこで、すでに原因系の変数と結果系の変数で群分けが可能である場合あらかじめ群を分けて解析をすることは有効な手段である。つまりグラフィカル連鎖モデリングとは変数をいくつかの群に分けて、原因系の方から第１群、第２群…とし、順番に解析をしていく解析方法である。下群より順番にグラフィカルモデリングをしていき、群同士の線に対しては矢線と考え数字の低い群から高い群へと矢線を結び、同じ群内の変数同士の線に対しては因果の向きはつけずに線とする。(宮川 [3]) つまり

1. 下位層をGMをし、線を消していく。
2. 一つ上の層を加えGMをする。その際下位層内の偏相関係数が 0 に近くても線は消さずに残しておく。
3. 求めた偏相関係数行列の下位層部分を手順 1 で求めた偏相関行列に置き換え作業を繰り返す。

この作業を行うことにより連鎖独立グラフを得る。図 1 のモデルにおいて１群={A,B}、２群={C}、３群={D,E} とした場合、連鎖独立グラフは図 3 となる。

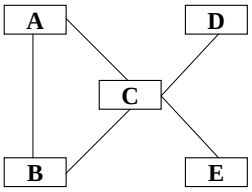

図 2　無向独立グラフの例

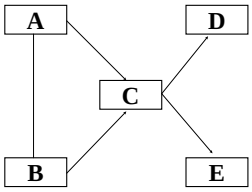

図 3　連鎖独立グラフの例

4 有向グラフの作成

グラフィカルモデリングやグラフィカル連鎖モデリングを用いて作成したモデルよりパス解析を行うのだが、ほとんどの場合全てに因果の向きが決まっているわけではない。パス解析を行うためには全ての因果の向きを決める必要がある。因果の向きを決めるに当たって、因果合流をする説明変数同士は独立にはならないということや変数の意味を考慮しながら因果の向きを決めていくのだが、最終的にいくつか候補があがる場合がある。このような場合は赤池情報量基準（AIC）を用いると良い。AIC とは、尤度で定義された 2 個以上の統計モデルのよさを数値で表すものであり、標本モデルと推定モデルが似ているほど期待平均対数尤度が高くなる性質があり

$$AIC = \chi^2 - 2df \tag{4}$$

で求めることができる。よって候補にあがったモデルの AIC を求めることにより一つの選ぶことができる。

5 プログラム

本研究ではパス解析を行うプログラムとグラフィカル連鎖モデリングを行うプログラムを作成した。最尤推定法のアルゴリズムにおいて、初期値を決めその値から最尤解へ収束するまで N 次元グラフ上を動いていくのだが、収束速度を早くさせるために、まず回帰分析を行い回帰係数を初期値とすることにより格段に速くなるよう工夫した。また、初期値からグラフを移動していく際に、移動幅に上限をつけることによって最初の動き出しの幅を大きくすることが可能となり収束速度が多少速くなるよう工夫した。

6 おわりに

本研究の最終目標であった「全自動解析」を作成する所までには至らなかった。今回実際にいろいろな解析し数え切れないほどパス図を作ってきたが、グラフィカル連鎖モデリングの群分けにおいて、ある程度分けられていれば変数の意味を取り入れなくてもある程度当てはまりの良い有向グラフを作成することができることが分かった。もちろん全てのデータに有効というわけではないが、この考えをプログラミングすれば無向グラフや連鎖独立グラフから有向グラフに変換することにより「全自動解析」に近いものが作れるのではないかと思う。

すでにグラフィカル連鎖モデリングを自動で解析するプログラムはある程度完成しているのだが、まだ思考ルーチンの改良の余地は十分にあると思われる。よって、もっといろいろなパターンのモデルを解析し、より最適なモデルを導くことができるプログラムを作成しいき、引き続き「全自動解析」を目標に今後もプログラムの改善に取り組んでいきたい。

参考文献

[1] 豊田秀樹：共分散構造分析「入門偏」, 朝倉書店,1998.
[2] 豊田秀樹：共分散構造分析「応用偏」, 朝倉書店,1998.
[3] 宮川雅巳：グラフィカルモデリング, 朝倉書店,1997.
[4] 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.