探索的パス解析の改良

M2021SS005 田野上正和
指導教員：松田眞一

1　はじめに

データサイエンスという分野が広く知られる今日，多変量データを解析する手法は様々に存在する．変数の有意性を推定する手法として頻繁に重回帰分析が挙げられるが，これは単独の目的変数に対する効果を観察することのみに留まっている．これに対して変数同士の因果を推定できる手法であるパス解析は特に有力な解析手段と考えられる．本研究では吉岡・松田 [11] による探索的パス解析を改良しつつ，議論の余地が残されていた点について検討する．

2　パス解析

2.1　パス解析とは

パス解析とは，Wright[10] が提案した解析手法である。これは連続型多変量データについて変数間の相関を非巡回有向グラフで表現される因果モデルによって推定を行うものである．

2.2　情報量規準

本研究では AIC と BIC という二種類の情報量規準を用いる．あるモデルと標本データから計算される情報量規準の値が低いほど良いモデルとされる．これら情報量規準の絶対的な数値に意味はなく，モデルの相対的な比較のために用いられる．以下の式で $\log(L)$ は対数尤度関数, k はパラメータ数, n はサンプルサイズを表す．

AIC（Akaike's Information Criterion）

Akaike [1] が提案した指標で，対数尤度とパラメータ数に基づいてモデルを評価する情報量規準である．

$$AIC = -2\log(L) + 2k \tag{1}$$

BIC（Bayesian Information Criterion）

Schwarz [7] がベイズ理論を基に導出した情報量規準である （赤池 [2] 参照）．

$$BIC = -2\log(L) + k\log(n) \tag{2}$$

本研究では，これらの情報量規準を小島 [4] に基づき修正した算出方法である，榊原 [6] の pas 関数を改良した吉岡・松田 [11] のプログラムによって求める．

2.3　AIC 最小化法

AIC 最小化法とは，良いモデルは最小の AIC を与えるという前提の下に探索を行う手法である（赤池 [2] 参照）．なお，赤池 [2] ではモデルの “良さ” と “正しさ” は異なることが指摘されている．本研究では，これに基づいてモデル探索を行う．

2.4　既存のパス解析における課題

通常，パス解析及び共分散構造分析を行う際は初めに前提となる因果仮説を立て，これを解析にかけながら妥当性を検証・修正していくといった手法が取られている．

ところがこの問題点として，データを解析する際に都度解析者による仮説モデルが必要であるのは体系的な統計解析とは程遠いものであることや，仮説モデルとは全く異なる構造の最適モデルを見逃すケースが十分に想定されることが挙げられる．これに対して吉岡・松田 [11] では仮説モデルを必要としない解析のために OR 手法の一つである山登り法と構造学習を採用し，乱択アルゴリズムによって近似最適解を探索する手法を提案している．本研究では吉岡・松田 [11] の研究を引継ぎ，より良い探索のために改良を施していく．

3　2 つのモデルの有意差検定

異なる因果モデルを比較するために，“2 つのモデルに差が無い” という帰無仮説を検定するための手法を紹介する．

比較する 2 つのモデルが包含関係にあるときは尤度比検定を，包含関係にないときは Linhart の AIC 有意差検定を適用する (下平 [8] 参照).

3.1　尤度比検定

尤度比検定では “比較する 2 つのモデルの χ^2 値の差” を χ^2 統計量，“両モデルの自由度の差” を自由度として χ^2 検定を行う (小島 [4] 参照).

3.2　Linhart の AIC 有意差検定

AIC の差を推定した標準偏差で割った統計量

$$z = \frac{\Delta AIC}{\sqrt{\hat{\mathrm{var}}(\Delta AIC)}} \tag{3}$$

が標準正規分布に従うとして両側検定を行う (下平 [8] 参照).

4　多重検定法

この章の内容は松田 [5] を参照する．

4.1　多重検定法とは

多重検定法とは，検定を同時に複数実施する際に FWER や FDR を制御することで全体で得られる結論の誤りを抑える検定方法である．本研究の問題で設定する仮説の数は非常に多く，FDR を制御することが妥当と考えられる．

4.2　FDR とは

m 個中 m_0 個が真の帰無仮説であるような仮説を同時に検定する問題を考える．真の帰無仮説と偽の帰無仮説を

採択・棄却する未知の確率変数を表 1 のようにそれぞれ U,V,T,S とするとき，$Q = \frac{V}{R}(R = 0$ のとき $Q = 0$ とする) が FDR の定義である．本来は Q を制御したいが，これは決して知りえない値であるために多重検定法ではその期待値 Q_e を制御することとしている．

表 1　m 個の帰無仮説に対する検定結果の分割表

	検定		
	採択	棄却	計
真の帰無仮説	U	V	m_0
偽の帰無仮説	T	S	$m - m_0$
計	$m-$R	R	m

4.3　BH 法

本研究ではモデルを比較的多数棄却することで棄却されないモデル集合を明らかにしたい狙いがあるため，FDR を制御する BH 法を採用する．BH 法の手順については松田 [5] を参照する．

5　モデル信頼集合作成手順

本研究では下平 [8] に倣って，“最良モデルと差がある” とは言えないモデルの集合をモデル信頼集合と呼称する．以下に棚橋 [9] を参考とした，本研究におけるモデル信頼集合作成の手順を示す．なお，基準となるモデルは最良モデル・最適モデルの 2 通りで考える．

モデル信頼集合作成手順

1. AIC 規準で探索を h 回行い，AIC 値が昇順に並ぶ $M_1, M_2, ..., M_r$ の r 通りのモデルを得る．

2a. 基準を最適モデルとする場合
   （a）最適モデルを全探索によって求め，M_0 とする．
   （b）帰無仮説 T_i を “M_0 と M_i のモデルの適合度に差がない”$(i = 1, 2, ..., r)$ と定める．

2b. 基準を最良モデルとする場合
   （a）帰無仮説 T_i を “M_1 と M_i のモデルの適合度に差がない”$(i = 2, 3, ..., r)$ とする．

3. T_i は比較するモデルが階層関係にあれば尤度比検定，階層関係になければ Linhart の AIC 有意差検定として実施し，対応する p 値 P_i を求める．

4. 得られた P_i に基準 $q*$ で BH 法を適用し，棄却されなかった帰無仮説に対応するモデル群と基準モデルを併せてモデル信頼集合とする．

6　データについて

本研究で使用するデータについて説明する．データは吉岡・松田 [11] で用いられたものと同様のものである．

消費者データ

9 変数，サンプルサイズ 47 のデータ．2014 年全国消費実態調査において都道府県別に集計されたデータである．変数の項目は月消費支出，食費，住居費，年間収入，貯蓄現在高，負債現在高，世帯数分布，世帯主の年齢，持ち家率 である．

6 変数消費者データ

6 変数，サンプルサイズ 47 のデータ．上記の消費者データから食費，年間収入，負債現在高の 3 変数を除いた 6 変数の消費者データである．

打者データ

11 変数，サンプルサイズ 43 のデータ．2003 年プロ野球選手のうち規定打席に到達した打者の成績データである．変数の項目は, 試合数，打数，得点，安打，本塁打，打点，盗塁，四球，年齢，年俸，優勝である．変数の “優勝”は, リーグ優勝のダミー変数である．

マーケットデータ

11 変数，サンプルサイズ 2216 のデータ．kaggle[3] で公開されている，Market Analytics データにおけるデータである．変数の項目は year_Birth，Income，kids，Recency，NumDealsPurchases, NumWebPurchases, NumCatalog-Purchases, NumStorePurchases, NumWebVisitsMonth，AcceptedCmp1，Response である．

7　先行研究で作成されたプログラムの改良

榊原 [6]，吉岡・松田 [11] で作成された R プログラムを高速化するため，以下の方針に沿って改修を行った．

(I) 計算量の少ない方法で計算を行う．

(II) 同じ計算結果を再利用する．

(III) R が得意とするベクトル・行列演算の記述を用いる．

(IV) 結果が変わらない範囲で，精密な計算を省略する．

改修前後の実行時間比較を表 2 に示す．行った探索は山登り法，100 回反復である．サンプルサイズが 2000 を超えるマーケットデータでは 2 割程度の改善に留まったものの，サンプルサイズが 50 未満の消費者データ，打者データでは 6 割程度も短縮される結果となった．

表 2　改修前後の実行時間比較（単位：秒）

対象データ	改修前	改修後
消費者データ	76.8	31.9
打者データ	153.6	57.6
マーケットデータ	744.2	592.7

8　減少法の提案

本章では辺に基づく減少法を提案し，山登り法および bnlearn との比較を行う．この手法では最も有向辺の多い

グラフを初期状態とすることで AIC 値の良い複雑なグラフが得られること，吟味する辺の数が半減することで実行時間の短縮が期待される．以下に減少法の手順を示す．

減少法の手順

1. 辺を最大数持つ有向非巡回グラフに対応するパス行列 A を初期解とする．
2. パス行列 A の有向辺について，吟味する辺の順番を無作為に生成する．
3. 手順 2 で生成された順に辺を吟味する．
   3.1. 暫定として辺を除いたモデルの AIC 値を計算する．
   3.2. 元のモデルに比べて AIC 値が減少する又は変化しないならば，その辺を削除する．
4. 全ての辺について吟味し終えたとき，その時点で残っているパス行列を解とする．

減少法の初期解となる辺を最大数持つ有向非巡回グラフを生成し，これを保証するため，二つの命題を提示する．

命題 1

あるパス行列 A が対角成分が全て 0 である上三角行列のとき，パス行列 A は有向非巡回グラフを示す．

命題 2

あるパス行列 A が辺を最大数持つ有向非巡回グラフに対応しているとき，任意に変数を入れ替えることでグラフに対応するパス行列を上三角行列にできる．

命題 1 証明

パス行列 A が巡回有向グラフを示すと仮定する．巡回路を構成するノード集合のうち添字が最小であるノードを x_i とする．そのとき，x_i は巡回路に含まれるノードであるため，x_i に入る有向辺が必ず存在する．その有向辺の有向元のノードを x_j とすると，$i < j$ の関係となるため，有向辺に対応するパス行列の成分は下三角行列に含まれ，仮定と矛盾する．よって命題が証明された．□

命題 2 証明

ノード $x_1, ..., x_n$ からなる，辺が最大数ある有向非巡回グラフに対応するパス行列 A を考える．順列集合 $S = \{x_1\}$，$i = 2$ とする．このとき S に対応するパス行列を A' とすると，A' は自明に上三角行列である．上三角行列に対応している順列集合 S に新たに x_i のノードを加えることを考える．ノード $x_1, ..., x_n$ からなるグラフは非巡回であるので，順列集合 S とノード x_i のみからなるグラフも非巡回である．ここで，集合 S のうち，x_i に向くノードの順列集合を S_1，x_i から向くノードの順列集合を S_2 とする（S_1, S_2 は空集合でもよい）．S_1, x_i, S_2 という並びになるよう順列集合 S に x_i を加える．S_1, S_2 の要素数をそれぞれ n_{S_1}, n_{S_2} とすると，S に対応するパス行列 A' の $1, ..., n_{S_1}$ 行 $n_{S_1}+1$ 列成分と $n_{S_1}+1$ 行 $n_{S_1}+2, ..., n_{S_1}+n_{S_2}+1$ 列成分に 1 が立ち，上三角行列が保たれる．数学的帰納法により，$i = n$ のときに構成される順序集合 S の並びに対応するようパス行列 A を並び替えたパス行列 A' は上三角行列となり，命題が証明された．□

8.1 事例解析

消費者データに減少法を適用し，山登り法および bnlearn との結果を比較する．山登り法と減少法はそれぞれ 10 回，100 回，1000 回反復で実行した．BIC 規準でモデルを探索した場合の最良モデルを表 3 に示す．seed 値を 1 から 100 まで変化させて 100 回繰り返し探索を行い，AIC 規準の各手法で得られた最良 AIC 値の箱ひげ図を図 1 に示す．

表 3 消費者データに対する各手法比較（BIC 規準）

探索方法	反復回数	BIC	edges	実行時間 (秒)
山登り法	10	−45.34	10	2.9
山登り法	100	−53.14	15	30.9
山登り法	1000	−54.85	11	308.4
減少法	10	−54.30	15	1.5
減少法	100	−58.66	16	16.1
減少法	1000	−60.07	13	166.5
bnlearn	1000	−57.40	13	0.3

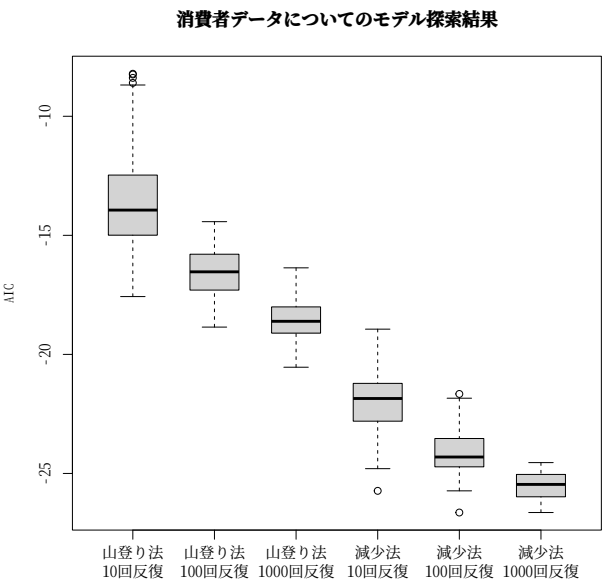

図 1 探索手法及び反復回数別 AIC 値の箱ひげ図

図 1 から，実行時間・最良 AIC 値を比較するといずれも減少法がより良い値を示している．更に，減少法は 10 回反復で山登り法 1000 回反復よりも良い解を得られていることが分かる．また表 3 から，減少法を 100 回以上反復することで bnlearn を上回る解を発見することができた．最良モデルの構造については減少法と bnlearn の最良モデル

（図 3, 4）で共に食費が始点となっており，食費から様々な変数を推定するモデルは妥当といえる．最も BIC 値の良かった減少法による最良モデル（図 3）では，山登り法と bnlearn（図 2，4）で共に終点に選ばれた貯蓄現在高に全く有向辺が繋がれず，興味深い結果となった．

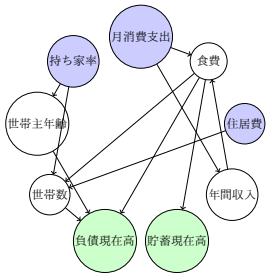

図 2 消費者データ,BIC 規準山登り法の最良モデル（BIC=−54.85）

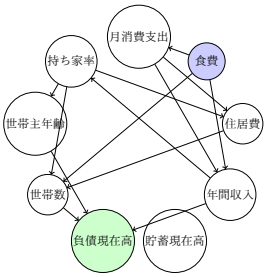

図 3 消費者データ,BIC 規準減少法の最良モデル（BIC=−60.07）

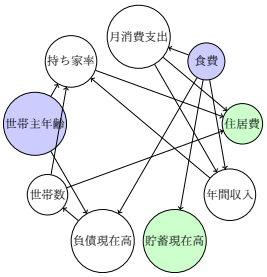

図 4 消費者データ,bnlearn の最良モデル（BIC=−57.40）

9 モデル信頼集合の作成

各データについて AIC 規準の減少法で 1000 回探索して得られたモデル群を検定し，モデル信頼集合を定める．BH 法による多重検定で棄却されなかったモデル群をモデル信頼集合とする．BH 法で用いる $q*$ の値は 0.05 とした．

9.1 6 変数消費者データに対する作成結果

6 変数消費者データに AIC 規準で減少法を 1000 回反復して得られた 259 通りのモデルから信頼集合を作成した結果，12 のモデルから構成された．これに含まれるモデルの順位は 1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 18, 35 であった．上位 10 モデルについての検定結果を表 4 に示す．最良モデルは貯蓄現在高が終点になっていたが，月消費支出や世帯数といった変数が終点となる良いモデルについても “最良モデルと差がある”とは言えないという結論が得られた．

10 まとめ

本研究で提案した減少法をモデル探索に用い，合わせてプログラム改修を行うことで，短時間でよりデータに適合した最良モデルを探索できた．また，十分な反復回数の試行を行うことで bnlearn を上回るモデルも得られた．

モデル信頼集合の作成手順をモデル比較検定と多重比較法によって示した．減少法によって得られたモデル群からモデル信頼集合を作成した結果，様々な構造のモデルが “最良モデルと差がある”とはいえない結果となった．

表 4 6 変数消費者データ減少法上位 10 モデル検定結果

id	AIC	dup	edges	sd	p 値	非棄却
32	−8.00	6	9	−	−	−
125	−8.00	4	9	0	1.000	○
144	−8.00	5	9	0	1.000	○
25	−7.95	7	9	0.60	0.931	○
127	−7.95	6	9	0.60	0.931	○
206	−7.95	1	9	0.60	0.931	○
67	−7.69	4	9	0.64	0.639	○
105	−7.31	7	9	0.64	0.639	○
365	−7.31	4	9	0.45	0.128	○
28	−7.04	5	10	0.45	0.034	×

11 おわりに

本研究の目的は吉岡・松田 [11] における探索手法を改良すること，そしてモデル信頼集合の作成手順を示すことであった．一つ目の目的は減少法の提案およびプログラム改修によって達成され，二つ目の目的も実現できたといえる．

一方，未解決問題として残っている点も多く，これらについては後続研究への課題としたい．

参考文献

[1] Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models, *Biometrika*, **60**(2), 255-265.

[2] 赤池弘次・甘利俊一・北川源四郎・樺島祥介・下平英寿 (2007).『赤池情報量規準 AIC』. 共立出版.

[3] kaggle : https://www.kaggle.com/

[4] 小島隆矢 (2003).『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.

[5] 松田眞一 (2008). FDR の概説とそれを制御する多重検定法の比較. 計量生物学, Vol. **29**, No. 2, pp.125-139.

[6] 榊原浩晃 (2005).『R によるパス解析の実現とその応用』. 南山大学数理情報学部数理学科卒業論文.

[7] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461-464.

[8] 下平英寿・久保川達也 (2004).『モデル選択 -予測・検定・推定の交差点-』. 岩波書店.

[9] 棚橋昌也 (2008).『AIC の多重比較法によるモデル集合の研究』. 南山大学数理情報学部数理学科修士論文.

[10] Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics*, **5**(3), 161-215.

[11] 吉岡祐輔・松田眞一 (2021). 探索的パス解析に関する研究. 南山大学紀要『アカデミア』理工学編, Vol.**22**, pp.106-123 .