探索的パス解析に関する研究

M2019SS701 吉岡裕輔
指導教員：松田眞一

1　はじめに

得られたデータの因果関係をモデル化し, 視覚的に把握する方法として, 有向非巡回グラフを用いることがある. 有向非巡回グラフとは, 矢印とノードで構成された有向グラフで閉路を持たないグラフである. 図 1 に例を示す. 現在, 有向非巡回グラフの構造を推定する手法として, ベイジアンネットワークを利用することが一般的である. しかし, ベイジアンネットワークは, ベイズ理論を基調としているので, 離散型変数のみしか扱うことができない. そのため, 連続型変数の場合は, 情報量の損失を押さえつつ離散化しなければならず, モデルの推定は容易ではない.

これまで, データが連続量の場合は, グラフィカルモデリングや共分散構造分析が利用されてきた. しかし, これらから有向非巡回を自動で作成することはできない. そこで, 本研究では連続型変数を離散化せず, R での榊原 [10] のパス解析関数と山登り法を用いて, 自動で有向非巡回グラフ構造を構築するプログラムを R で作成し, その性質を研究していく.

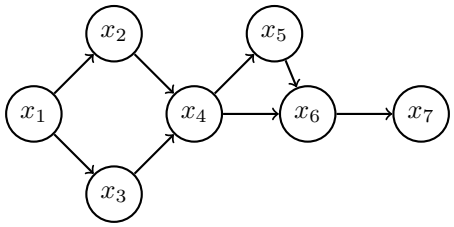

図 1: 有向非巡回グラフ例

2　パス解析

2.1　パス解析とは

パス解析とは, Wright [13] [14] が考案した, 多変数の因果関係を線形相関の計算からパス係数を求め, パス図を作成する手法である. コンピュータ性能が向上した現代では, 共分散構造分析 (SEM) の下位モデルとしての位置づけである.

2.2　パス図の適合度評価

パス解析では, 作成したモデルを評価するために適合度指標を用いる. 指標は 50 種類以上あるが, 本研究は, 榊原 [10] が使用した指標に BIC を加えたものを使用する.

情報量規準

得られたデータから構築されたモデルの当てはまりの良さを評価するために使われる. AIC, BIC 共に数値が低い方がより良いモデルとされている.
本研究では, ベイジアンネットワークとの比較を行うため, BIC を基本の情報量規準として用いる.

以下, log(L) を対数尤度関数, k をパラメータ数, n を標本の大きさとする.

AIC（Akaike's Information Criterion）
Akaike [1] によって提案された. モデルの評価を最尤法に基づいて行う情報量規準である.

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

BIC(Bayesian Information Criterion)
Schwarz [11] がベイズ理論を基調として導出した. また, Schwarz [11] は AIC はモデル選択で真のモデルを選ばないとし主張している. (赤池 [2] 参照)

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

3　山登り法

山登り法は, 局所探索法 (local search) として, Johnson *et al.* [6] によって提案された. その後に Heckerman *et al.* [5] が, 評価関数を使ってベイジアンネットワークのスコアを計算し, 構造を推定することに応用した. 山登り法の問題点として, 局所的最大値に陥ってしまうことが挙げられる. これに対し, Heckerman *et al.* [5] は, 局所探索法が潜在的に持つ問題点の回避方法について以下のように述べている. 「局所的最大値を回避する方法は, 反復局所探索法 (iterated local search) と焼きなまし法 (simulated annealing) が挙げられる. 反復局所探索法においては, 我々は, 局所探索法が局所最大値に達するまで適用する. 次に, 現在の構造を無作為に摂動させ, 操作可能な回数の反復で過程を繰り返す. すべての段階で, 一番構造が良かった構造を保持する.」

4　ベイジアンネットワーク

4.1　ベイジアンネットワークとは

ベイジアンネットワークとは, 頂点が変数を表し, 辺が繋がった頂点間の直接的な因果関係の影響力を示す. これらの影響力の大きさが条件付き確率によって定量化された有向非巡回グラフである. (Pearl [9] 参照)
Pearl [9] 以前は, 確率を使用したネットワークは, 複数の呼び名があったが, Pearl [9] 以降はベイジアンネットワークと一般化された. (Eugene [3] 参照)

4.2　bnlearn について

bnlearn は, Scutari [12] を基に, R Development Core Team よって作成された R パッケージであり, 複数のアルゴリズムの使用が可能である. 離散型変数からは, ベイジアンネットワークの構造学習を行うことができる. また, 連

続型変数では, Geiger and Herckerman [4] と Neapolitan [8] の論文を基にして, データ構造が正規分布に近似できるときに限り, 連続変量における条件付き独立性を使用して, ガウシアンネットワークを構成することができる.

5　プログラミング

本研究で作成した山登り法とパス解析の手法を組み合わせ, モデルを複数探索するプログラムと探索過程を 2 段階に分ける構造学習プログラムについて述べる.

本研究での山登り法探索プログラム

関数内の処理の手順は以下の通りである.

本研究での山登り法の手順

Step1: 空グラフを作成する.

Step2: 任意に選択する辺の順番を決める.

Step3: 選択された辺に対して, 任意の辺を追加, 反転, 削除を行った際, 有向グラフが巡回していないかを確認する. 巡回していなければ, BIC(AIC) 値を求め, 最も BIC(AIC) 値が改善するように有向非巡回グラフを構成する.

Step4: 追加, 反転, 削除を行い BIC(AIC) 値が同一の場合, 乱数を発生させ, 任意にグラフ構成を選択する.

Step5: すべての辺が選択された後, 有向非巡回グラフの BIC(AIC) 値を求め, 値が良いものから順にデータフレームへ格納して行く.

構造学習プログラム

山登り法によるランダム探索を 2 段階に分けて, 実行する手法である. 1 段階目で探索を行った結果から, 有効となる有向辺を学習し, 2 段階目で探索を行う際に初期の空グラフに有向辺を加えランダム探索を行う.

6　データについて

本研究の事例解析 (第 7 章) と シミュレーション (第 8 章) で使用するデータについて説明する.

カメラデータ　変数項目が小型軽量, 携帯容易, 操作用意, 総合満足であり, 標本サイズは 100 個. (小島 [7])

消費者データ　2014 年全国消費実態調査における都道府県ごとのデータ分析を行う. データの変数項目は, 食費, 居住費, 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.

消費者ミニデータ　上記の消費者データの変数を 2 つ削減したデータを使用する. データの変数項目は, 月消費支出, 食費, 住居費, 貯蓄現在高, 世帯数分布, 世帯主年齢, 持ち家率の 7 変数である.

7　事例解析

事例解析では, 消費者データの解析を行う. 目的変数, 説明変数を指定せず, 純粋なパス解析と山登り法により探索を行い, bnlearn との比較を行う. その後, 構造学習の手法と純粋なランダム山登り法の結果と比較, 検討をする.

7.1　解析結果

解析 1　パス解析に山登り法を適用させ, BIC 規準で 1000 回繰り返し, 得られた上位 15 モデルである. id は何度目の繰り返しかを示し, dup はモデルが重複した回数を示す.

表 1: 消費者データ 1000 回実施結果

	id	AIC	BIC	dup
1	768	−19.48	−56.48	1
2	279	−17.84	−54.85	1
3	145	−15.20	−54.05	1
4	883	−21.72	−53.17	1
5	234	−14.22	−53.08	1
6	475	−12.14	−52.84	1
7	48	−13.67	−52.52	1
8	998	−15.20	−52.19	1
9	865	−13.32	−52.17	1
10	166	−13.29	−52.14	1
11	652	−7.16	−51.56	1
12	730	−17.97	−51.28	1
13	85	−12.10	−50.95	1
14	132	−11.58	−50.44	1
15	694	−11.45	−50.31	1

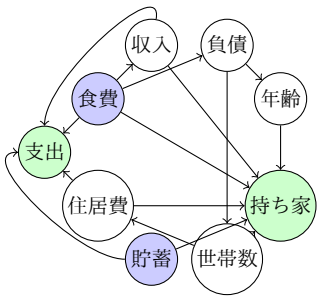

図 2: id_768 結果

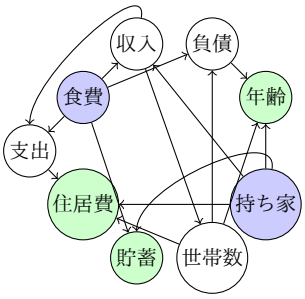

図 3: id_145 結果

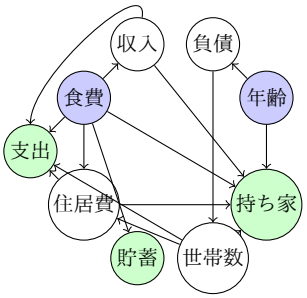

図 4: bnlearn 結果

パス解析から得られたモデルの中で, BIC 値が最も低いモデル, BIC 値が三番目に低いモデル, また, bnelarn から得られたモデルの有向非巡回グラフを

示した. また, bnlearn によって得られたモデルのパス解析結果は以下の通り.

表 2: bnlearn パス解析結果

chi^2 = 19.30	df = 20
AIC = −20.70	BIC = −57.70
AGFI = 0.82	GFI = 0.92
P-Value = 0.50	

表 3 は, 図 2, 図 3, 図 4 から始点, 終点となる頂点と得られた BIC 値をそれぞれ集計した結果である. 推定されたモデルの BIC 値の差はわずかであるが, 構造が異なるモデル群を探索した結果となった. よって, 単一の最良のモデルを求めるだけでなく, 複数のモデル構造を推定する手法が有効であると言える. また, bnlearn が推定したモデルは, 本研究手法から得られた最良モデルの BIC 値より, 約 1.3 低く良いモデルを推定した.

表 3: 始点終点の集計結果

	BIC	始点	終点
id_768	−56.48	食費, 貯蓄	支出, 持ち家
id_145	−54.85	食費, 持ち家	住居費, 貯蓄, 年齢
bnlearn	−57.70	食費, 年齢	支出, 貯蓄, 持ち家

解析 2 表 4 は構造学習において, 1 段階目の探索と 2 段階目の探索におけるパラメータの設定ごとの結果である. top は, 最良の AIC 値, medain は, 上位 15 モデルの中央値, mean は, 上位 15 モデルの平均値を示す. また, 純粋なランダム探索山登り法は, AIC 規準で山登り探索を 1000 回行い, 得られた上位 15 モデルから集計した結果である. 表 5 は, 構造学習を計 800 回を行った結果であり, 表中の wl_from は, 1 段階目の探索で得られた有効な有向辺の始点, wl_to は, 有効な有向辺の終点となる. このとき, 1 段階目の有効な有向辺の探索は, 100 回の繰り返しを行い, 2 段階目の探索は, 700 回行った.
表 6 は推定された上位 15 モデルにおける始点, 終点となった回数である.

表 4: 構造学習 実施結果 (AIC 規準)

	100_700	320_480	480_320	main
top	−18.91	−17.74	−17.58	−19.49
median	−17.42	−16.92	−16.25	−15.73
mean	−17.4	−16.98	−16.31	−16.1

構造学習の手法は, 純粋なランダム探索山登り法の結果と比較して, AIC 値が最良となるモデルを探索することはできなかったが, 中央値, 平均値からは, 安定したモデル群を探索できたと言える. 解析 1 より, 情報量規準の差がわずかでも作成するモデルが大きく異なることから, 解析時間を短縮し, 安定したモデル群を探索できる構造学習の手法は有用であった. また, 構造学習ではパラメータ設定により結果に差が生まれるため, 探索を行うパラメータの設定には注意する必要があると言える.

表 5: 100_700 実施結果 (AIC 規準)

id	AIC	BIC	dup	wl_from	wl_to
684	−18.91	−42.96	1	世帯数	負債
415	−17.74	−43.64	1	食費	貯蓄
522	−17.62	−36.12	1	世帯数	負債
220	−17.58	−43.48	1	世帯数	負債
384	−17.57	−39.77	1	世帯数	負債
627	−17.47	−41.53	1	食費	貯蓄
334	−17.42	−39.63	1	世帯数	負債
14	−17.42	−41.48	1	世帯数	負債
362	−17.23	−41.28	1	世帯数	負債
698	−17.16	−39.36	1	世帯数	負債
554	−17.12	−37.47	1	世帯数	負債
478	−17.12	−41.17	1	世帯数	負債
444	−16.92	−46.52	1	世帯数	負債
144	−16.89	−39.09	1	世帯数	負債
594	−16.84	−39.04	1	世帯数	負債

表 6: 構造学習が推定したモデル群の始点と終点

	支出	食費	住居費	収入	貯蓄	負債	世帯数	世帯主	持ち家
始点	14	15	11	15	0	0	15	0	6
終点	0	0	3	0	7	7	0	2	6

表 6 から, 持ち家と世帯主以外の頂点は, 始点と終点になりやすい頂点にそれぞれ分かれていることがわかる. しかし, 持ち家は始点と終点となったモデルが同数あり, 変数が振動していることがわかる. ここから, bnlearn のようにモデルを単一に推定する手法が危険であって, モデルを複数探索する本手法が有効であると言える.

8 シミュレーション

第 7 章の事例解析では, 実データを使用して, より良い評価値を得られるようにモデル構造を探索的に求めた. 本章では, カメラデータと消費者ミニデータでモデルを定義し, 真のモデル構造からデータを与えて解析した際, 再現性がどれほど確保されるのかを検証を行い明らかにする. なお, ここでは紙面の都合上カメラデータの結果のみを示す.

8.1 シミュレーションの説明

シミュレーションを実行する際, 真の構造とする有向非巡回グラフを選択し, そのパス行列とパス係数行列を使用して, データ生成を行う.

比較手法について

真のモデルとシミュレーションで得られたモデルとを比較して, (i) 必要な有向辺がない数, (ii) 不必要な有向辺の数, (iii) 真のモデルの有向辺と一致した有向辺の数, (iv) 真のモデルとの距離 ((i) と (ii) の和) を再現性の評価として使用する.

シミュレーション手順の詳細について

シミュレーションの手順

Step1: 真のモデルを指定し, データ生成を行う.

Step2: 純粋なランダム探索山登り法と構造学習を実行し, シミュレーションを行う.

Step3: パラメータ設定は純粋なランダム探索山登り法では, 探索回数を 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 に設定し, 各 10 回シミュレーションを行う. 構造学習の探索回数は, 純粋なランダム探索山登り法の探索回数の 8 割とする. 構造学習の実行パラメータは, 1 段階目の探索は, 探索回数の 6 割, 2 段階目の探索を残りの 4 割とした.

Step4: 真のモデルとシミュレーションによって得られたモデルとの比較を行う.

8.2 シミュレーション結果

図 5, 図 6 は, カメラデータにおけるシミュレーション結果である. これらの図は, シミュレーションの各回ごとに上位 10 個のモデルの表を作り, 距離の重み付き平均から作成した箱ひげ図である. 重みはモデルの重複を使用した.

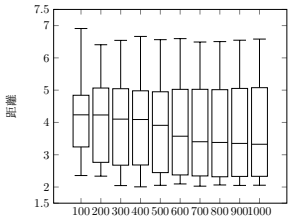

図 5: 純粋な山登り法結果

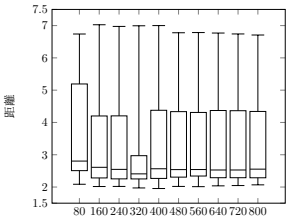

図 6: 構造学習結果

純粋な山登り法の結果について 探索結果の安定性を示す中央値が改善していく様子が見られたが, 探索回数が 700 回あたりから解が飽和している様子が見られる. また, 全体の探索結果のばらつきは改善しなかったが, 区間ごとのデータのばらつきでは, 中央値から第一四分位までの間にデータが集中していく様子が見られ若干改善している.

構造学習の結果について 探索回数が 100 回あたりで既に解が飽和していた. また, 中央値は常に安定しており, 安定した探索結果が得られた. そして, データのばらつきは中央値から第一四分位までの間にデータが集中している.

シミュレーション結果の総括 これらの結果より, 構造学習は純粋な山登り法の結果に比べ, 安定したより高い再現性の結果を探索できたと言える.

9 まとめ

パス解析においてモデルを構築する手段がこれまで確立しておらず, 経験則, 連続量の離散化から条件付き独立性の使用, bnlearn などからモデル探索が行われてきた. 本研究ではパス解析と山登り法を使用して, 連続量から複数のモデル探索を可能とし, 探索時間の短縮と安定したモデル群の探索のため構造学習の手法を提案した. そして, 事例解析, シミュレーションの結果から構造学習の有効性を示すことができた. また, 本研究手法は表 6 のように揺らぎのあるデータに対して, 複数のモデル構造を探索できるため特に有用であるが, bnlearn は, 単一のモデルしか探索できないため, データの因果関係を十分に把握することができないと言える.

10 おわりに

本研究ではパス解析と山登り法を改良することにより, 複数のモデルを探索する可能とすることができた.

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 (1996). AIC と MDL と BIC, 『オペレーションズ・リサーチ』, **41**(7), 375-378.

[3] Eugene, C. (1991). Bayesian networks without tears, *AI Magazine*, **12**(4), 50-63.

[4] Geiger, D. and Heckerman, D. (1994). Learning Gaussian Networks, *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, 235-243.

[5] Heckerman, D., Geiger, D. and Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning*, **20**, 197-243.

[6] Johnson, D. S., Papadimitriou, C. H. and Yannakakis, M. (1988). How Easy Is Local Search? *Journal of computer and system sciences*, **37**, 79-100.

[7] 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

[8] Neapolitan, R. E. (2003). *Learning Bayesian Networks*, Prentice Hall.

[9] Pearl, J. (1985). Bayesian Networks: a Model of Self-Activated Memory for Evidential Reasoning, *UCLA Computer Science Department Technical Report* 850021 (R-43).

[10] 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科学科卒業論文.

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

[12] Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package, *Journal of Statistical Software*, **35**(3), 1-22.

[13] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research*, **20**, 557-585.

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