独立成分分析によるベイジアンネットワークの構造推定

http://www1.m.jcnnet.jp/mabonki
mabonki@ka3.koalanet.ne.jp

1 ベイジアンネットワークの構造推定

データからベイジアンネットワークを推定できれば、データ項目間の関連が視覚的に把握でき
ベイジアンネットワークの確率伝播機能を使って因果を確率で計算できる。

データからベイジアンネットワークの構造推定する主な方法は以下の3つである。
1)制約ベース　：　2ノード間以外の他のノード群で条件独立を検定し、独立なら2点は非連結にする
　連結方向はD分離の法則を採用するが、全ての連結に方向が付くとは限らない。(4)

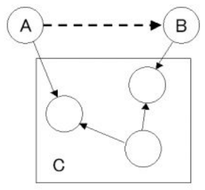


図1-1 条件独立による非連結化

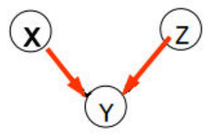


図1-2 D分離による連結方向付け

2)スコアベース：　全てのノードのリンクについて対数尤度を計算し、最大尤度を持つ構造を探索する。
　対数尤度指標には離散型モデルの最小記述長MDLが一般に使用される。(式2-1 参照)
　全ノードの連結と方向について対数尤度を計算する必要があるので、組合せの爆発が起こる。
　大規模なネットワークに適さない　　ノード数10個　4.2×10^{18} 組合せ　　(9)

3)無方向ネットワーク(GGMやグラフィカルLasso)を作成して、スコアベースで矢印を求める
　データからGGM(ガウシアン・グラフィカルモデル)やグラフィカルLassoで疎な無方向構造図を生成する。
　疎な連結の方向は対数尤度MDLが高い方にする。(表3-2参照)

上記の方法でタイタニック号の乗船名簿と生存のデータからベイジアンネットを構造推定した結果を示す。
当時の乗船名簿には氏名、性別、年齢、乗船等級、爵位、船室番号、船室階が記録されていた。(5)

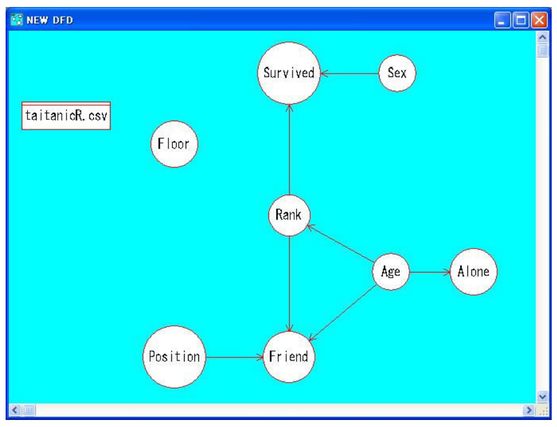


図1-3 条件独立による構造推定

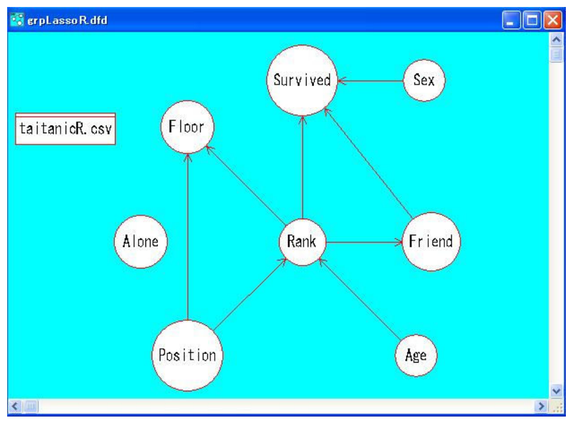


図1-4 グラフィカルLassoとMDLによる構造推定

Survived	生存有無
Sex	性別
Age	年齢
Friend	同船室
Position	爵位
Rank	乗船等級
Floor	船室階
alone	単独

構造推定されたベイジアンネットワークの見方

同船は沈没まで6時間かかっており、最初に女性を優先してボートに載せたので女性の生存率が高い。
生存率(女性：74% 男性：19%)
次に一等客を優先してボートに載せている。
両図とも性別と乗船等級ノードから生存有無ノードに矢印が向けられた構造が推定されている。

2 独立成分分析による因果推論

ノード間の連結方向を推定するには、条件を矢印元とする条件確率を使った2つのMDLを計算し
MDL値が高い方向に矢印を付ければよい。

生存有無		乗船等級
生存		1等
死亡		2等
		3等

MDL0 →　　← MDL1

$$MDL\,0 = -\log\left(\sum_{i=1}^{N} p(\text{乗船等級} \mid \text{生存有無}\ )\right) + \frac{k\ln(N)}{2}$$

$$MDL\,1 = -\log\left(\sum_{i=1}^{N} p(\text{生存有無} \mid \text{乗船等級}\ )\right) + \frac{k\ln(N)}{2}$$

式2-1

しかし、ノードのデータが実数の場合、離散型のMDLは使えないので、独立成分分析による因果推定する。

独立成分分析は、合成したデータを独立成分に分解することができる。

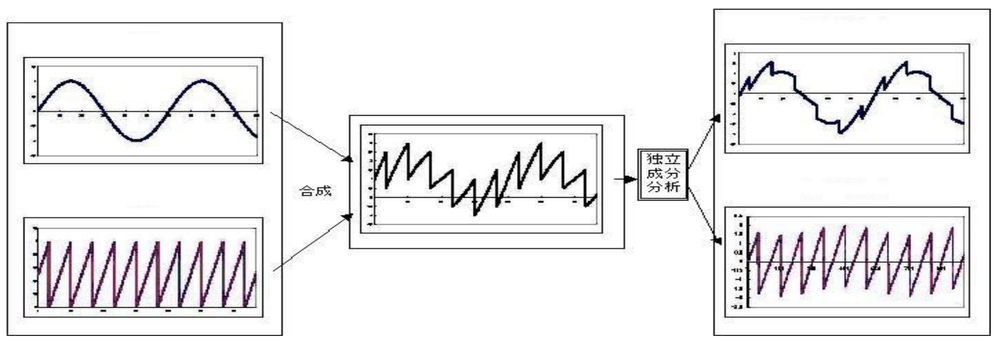


図2-1 独立成分分析による独立成分分解

両ノードが実数値の場合、独立に分解できる機能を使って因果分析を行う。

因果関係がある2系列のデータは互いに独立ではない。
これらを独立成分分析で分解すると、図2-2の様に独立な成分と分解行列が得られる。
得られた独立成分データは互いに独立なので、非独立の因果関係は分解行列に残るはずである。
この行列を加工すると、因果の強さと方向が判明する。

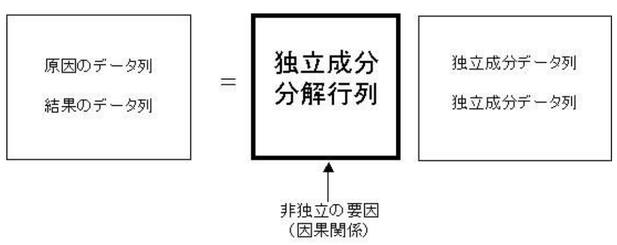


図2-2 独立成分分析による行列の分解

＊パイオニクス所属
　総合研究大学院　研究生　統計科学専攻

人工データで独立成分分析による因果分析を検証してみる。
図2-3の式で作成した(x,y)について独立成分分析を行うと、図2-4に独立した(x,y)が得られていることがわかる。

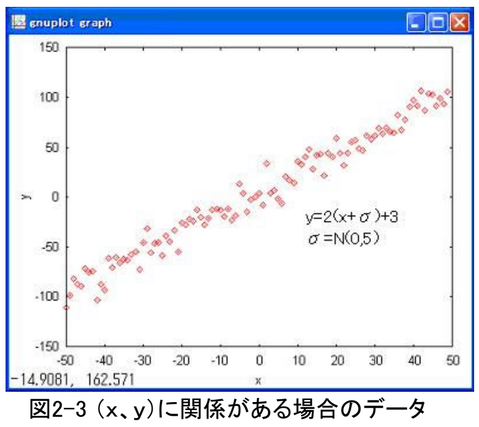


図2-3 (x、y)に関係がある場合のデータ

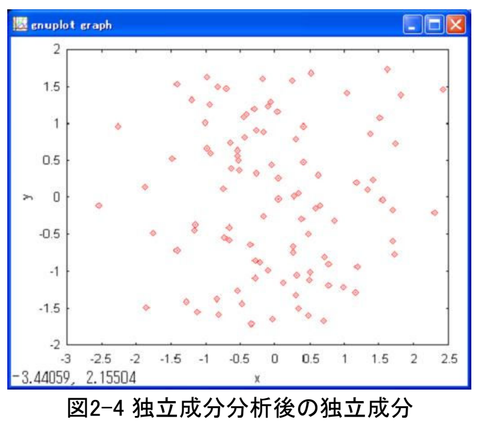


図2-4 独立成分分析後の独立成分

独立成分分解行列は図2-5の様に求まっている。これを下三角行列に変換する。
この行列によりxが原因である確率が(102.67　対　0.161)で得られる

Independent conponent matrix

data of x	0.2028	−1.033	ica of x
data of y	−0.1014	−0.001	ica of y

↓

lower Traiangle

0	0.161
−102.67	0

図2-5 独立成分分解行列の結果

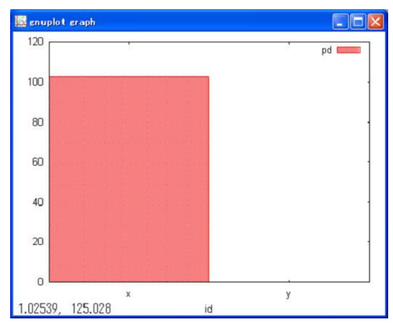


図2-6　xが原因である確率が(102.67 対 0.161)で得られる

3 独立成分分析を使ったベイジアンネットワークの構造推定の例

米国の銀行員の給与に人種の差別があるか調べるデータから構造推定する
（出典：新村秀一 PCによるデータ解析　付録2）

構造推定では、以下の手順で行なった。
1)　項目間の相互情報量のクロス表を求める。(表3-1)
2)　クロス表をグラフィカルLassoで疎なクロス表にする。(表3-2)
3)　変数をノードに割り当てる。
4)　疎なクロス表で、異なる変数間で値があれば、そのノード間を連結する。
5)　実数値のデータを持つノード間は独立成分分析で因果の方向を決定する
6)　離散のデータを持つデータ間はMDLで方向を決定する

表3-1　相互情報量

	jobcat	sex	minority	salbeg	salnow	age	edlevel	work
jobcat	1.371	0.112	0.045	0.155	0.187	0.138	0.062	0.125
sex	0.112	0.689	0.003	0.216	0.117	0.004	0.004	0.048
minority	0.045	0.003	0.526	0.002	0.016	0.021	0.001	0.019
salbeg	0.155	0.216	0.002	0.701	0.221	0.008	0.009	0.043
salnow	0.187	0.117	0.016	0.221	0.707	0.003	0.012	0.031
age	0.138	0.004	0.021	0.008	0.003	0.707	0.024	0.292
edlevel	0.062	0.004	0.001	0.009	0.012	0.024	0.365	0.018
work	0.125	0.048	0.019	0.043	0.031	0.292	0.018	0.872

表3-2　グラフィカルLasso

	jobcat	sex	minority	salbeg	salnow	age	edlevel	work
jobcat	1.431	0.058	0	0.066	0.074	0.074	0.018	0.048
sex	0.058	0.749	0	0.209	0	0	0	0
minority	0	0	0.586	0	0	0	0	0
salbeg	0.066	0.209	0	0.761	0.194	0	0	0
salnow	0.074	0	0	0.194	0.767	0	0	0
age	0.074	0	0	0	0	0.767	0	0.344
edlevel	0.018	0	0	0	0	0	0.425	0
work	0.048	0	0	0	0	0.344	0	0.932

ノード間が共に実数値を持つのは、(初任給 salbeg 現在給与　salnow)と(勤続年数 work と 年齢 age)の2箇所であった。

初任給と現在給与間で独立成分分析で因果を求めてみると、初任給が原因である確率は95%であった。

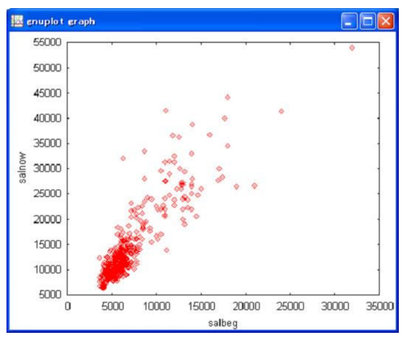


図3-1 初任給(salbeg) と 現在給与(salnow)との相関図

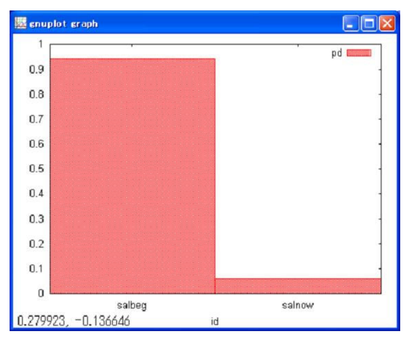
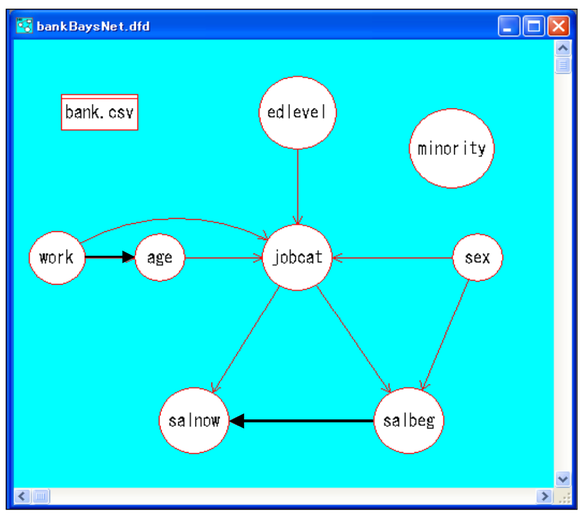


図3-2 初任給(salbeg) が 現在給与(salnow)の原因である確率

推定された構造図
少数民族(minority)は孤立していて、給与に関係ないことが示されている、
日本と異なって現在給与(salnow)と勤続年数(work)に因果関係はない。現在給与は職種(jobcat)と初任給(salbeg)で決まる。



jobcat	職種 (事務/警備/専門/技術/MBA/管理職)
mainority	少数民族
salnow	現在給与
salbeg	初任級
sex	性別
age	年齢
work	勤続年数
edlevel	教育期間

構造推定された給与ベイジアンネットワークの見方

①少数民族はそれ以外のノードと全く独立
②初任給は職種と性別で決まる。
③現在給与は職種と初任給で決まる。
④職種は教育期間、年齢、性別で決まる
⑤勤続年数から年齢が類推できる。

図3-3 米国銀行員の給与のベイジアンネット構造推定図

→　独立成分による因果推定

→　MDLによる因果推定

参考文献

(1) Kevin Murphy　§10 Direct Graphical Model「Machine Learning」
(2) 新村秀一　PCによるデータ解析
(3) 清水昌平　独立成分分析による線形逐次モデル
(4) Jie Cheng　Learning Baysian net work from data
(5) Y. Mitui　ベイジアンネットの学習アルゴリズム
(6) C.M.Bishop　§8 Graphical Model「Pettern Recognision and Machine Learning」
(7) 鹿島 久嗣　グラフィカルラッソ
(8) M.Nakai　TDPAベイジアンネット構造推定
(9) 植野真臣　ベイジアンネット