# データからのベイジアンネットワークの構造推定

# Inference Structure of Bayesian network from Data

総合研究大学院大学

Pionix Co., Ltd　The Graduate University for Advanced Studies

## Abstract：

There are 3 main methods of inference for Bayesian network from data. We compared their properties using actual example. Score base method by MDL(Minimum Description Length) can infer more appropriate structure than others. But this method occurs explosion of combination. So we proposed sparse network by significant link due to high MDL value and proved less problems for representation of structure by comparison with full link network.

## 1. はじめに

ベイジアンネットワークはノードと矢印で構成された有向グラフである。その重要な機能は確率伝播である。矢印で隣接されたノードの状態を条件とする条件付確率が、矢印を通じ次々伝播して全ノードの確率が得られる機能である[1]。データから有向ネットワークが自動的に生成できれば、データ項目間の影響関係が明らかになり、１つのノードの状態を変更することで全ノードの確率が変化し動的なモデルとして活用することができる。

データからベイジアンネットの構造を推定する主な方法として以下の３つがある[3]。

1) 制約ベース：条件付独立による連結
2) スコアベース：最小記述長(MDL)による連結
3) 統合ベース：無向グラフの有向化

## 2. 構造推定方法の比較

### 2.1 制約ベース

J. Pearl[2]の連結規則を適用した連結を行う。２ノード間が有意な連結でも、その他のノード群との条件付独立を検定し、非独立なら介在ノード群による遮断と見做しV字連結とする (D 分離の法則)。

$$I(A,B \mid C)=\sum_{a,b,c} P(a,b,c)\log\frac{P(a,b \mid c)}{P(a \mid c)P(b \mid c)}$$

I(A,B|C) ＞ ε なら条件付非独立とする。[3]

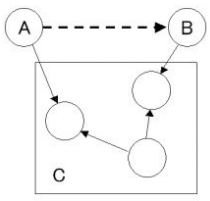



図2-1 条件付非独立の遮断

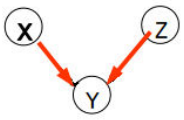
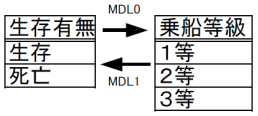



図2-2 V字連結

### 2..2 スコアベース

全ノード間の全リンク方向の組合せについてスコア即ち最小記述長(MDL:Minimum Description Length)を計算する。連結方向はMDLが高い方で連結する[4]。例えば２点間の連結では以下で計算する。

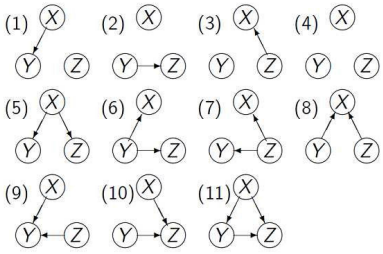



$$MDL\,0=-\log\left(\sum_{i=1}^{N} p(\text{乗船等級} \mid \text{生存有無})\right)+\frac{k\ln(N)}{2}$$

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

*連絡先　総合研究大学院大学　統計科学専攻　研究生
Email:mabonki0725@icloud.com

これは、一方を条件とした条件確率の対数尤度である。値が大きい方が高い説明力を示している。この場合、全ノード全方向で計算するので、組合せの爆発が発生する。例えばノードが3個でも11通りがある。




ノードが10個の場合 $4.2\times10^{18}$ となり、実用的な計算時間では収まらない。[5]

### 2.3 統合ベース

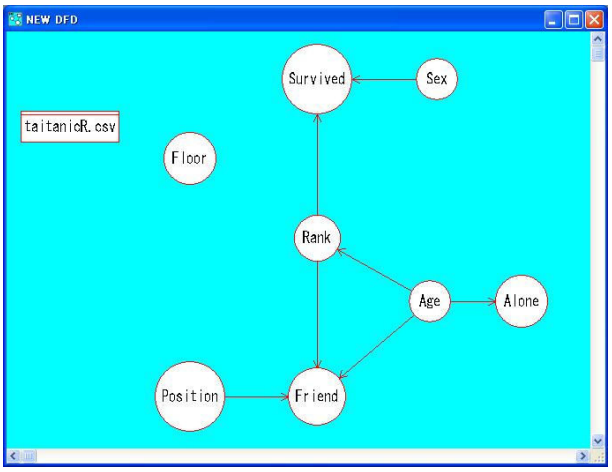
組合せの爆発を避けるためGGM(ガウシアングラフィカルモデル)[6][12]やグラフィカルLasso[7]で疎な無方向グラフを作成し、連結の向きをMDLで説明力の高い方にする。

## 3. 構造推定の適用例

1912年4月14日に発生したタイタニック号の乗船名簿と生存者のデータからベイジアンネットの構造推定をする。乗船名簿には1309名の氏名、性別、年齢、乗船等級、爵位、船室番号、船室階が記載されていた。[8]

### 3.1 制約ベース

2ノード間とその他のノード群とで条件付独立を検定する。閾値 $\varepsilon$ 以上は非独立と見做しV連結する。




| | |
|---|---|
| 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 |

次の条件付非独立が検出され。V字で連結されている。

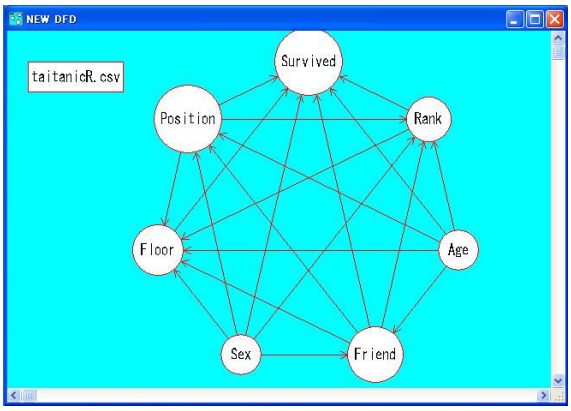
Sex と Rank が Survived で条件付非独立

Rank と Position が Friend で条件付非独立

Age の3連結は各2点間の有意な連結である。

### 3.2 スコアベース

このモデルには7ノードあり、全ノードと全リンク方向の組合せは1,138,779,265あった。全組合せのMDLを計算し最も尤度が高い構造推定を算出するまで計算時間はPCで7時間かかった。全ノードの矢印がSurvivedノードに向かっており、全ノードでSurvivedを説明するモデルとなっている。




### 3.3 統合ベース

疎な関係をGGMかグラフィカルLassoで求め、ノード間の方向はMDLの高い方で決定する。今回はノード間の相互情報量(次式)よりグラフィカルLassoで疎な関係を求めた。

$$I(A,B)=\sum_{a,b}P(a,b)\log\frac{P(a,b)}{P(a)P(b)}$$

表2-1 グラフィカルLassoの結果

| | Survived | Sex | Rank | Age | Friend | Position | Alone | Floor |
|---|---|---|---|---|---|---|---|---|
| Survived | 0.706 | 0.147 | 0.085 | 0 | 0.058 | 0 | 0 | 0 |
| Sex | 0.147 | 0.689 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rank | 0.085 | 0 | 1.038 | 0.138 | 0.040 | 0.211 | 0 | 0.153 |
| Age | 0 | 0 | 0.138 | 1.457 | 0 | 0 | 0 | 0 |
| Friend | 0.058 | 0 | 0.040 | 0 | 0.901 | 0 | 0 | 0 |
| Position | 0 | 0 | 0.211 | 0 | 0 | 0.735 | 0 | 0.657 |
| Alone | 0 | 0 | 0 | 0 | 0 | 0 | 0.703 | 0 |
| Floor | 0 | 0 | 0.153 | 0 | 0 | 0.657 | 0 | 0.979 |

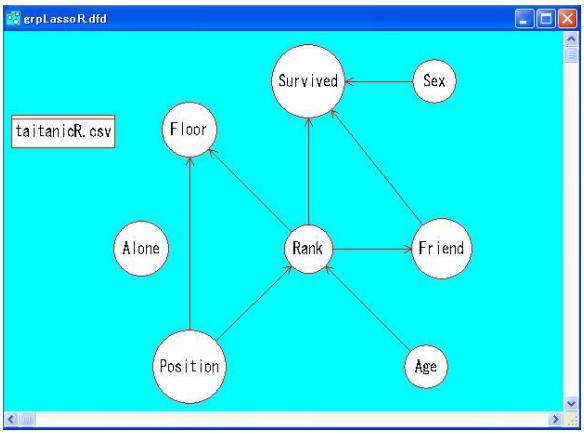
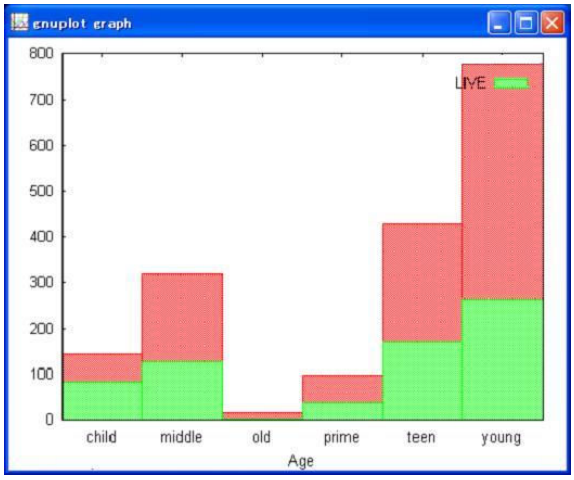



図3-3 疎なノード間をMDLで方向付けたモデル

### 3.4 考察

自動生成したベイジアンネットでは、性別と乗船等級が生存に関わっていることが共通している。これは以下の事実から説明できる。タイタニック号は座礁から沈没まで2時間かかっており、最初は秩序だった事故対策が採られた。即ち最初に女性を救命ボートに降ろし、次に1等船客を乗せている。女性の生存率は75%に対して男性は19%だった。疎なグラフでは年齢は生存に直接関係していない。年齢層別の生存率を下図でみると child 以外は全て半分以下で生存に関して年齢層に差がなく正しい反映と言える。




## 4. 有意な結線のみによる構造学習

制約ベースでは、条件付独立の検定と J.Pearl の連結規則での方向付けなので、直接的な影響関係でない。統合ベースは影響方向を無視した疎な無向グラフが基になっている。そこでスコアベースの組合せの爆発を回避する方法を検討する。

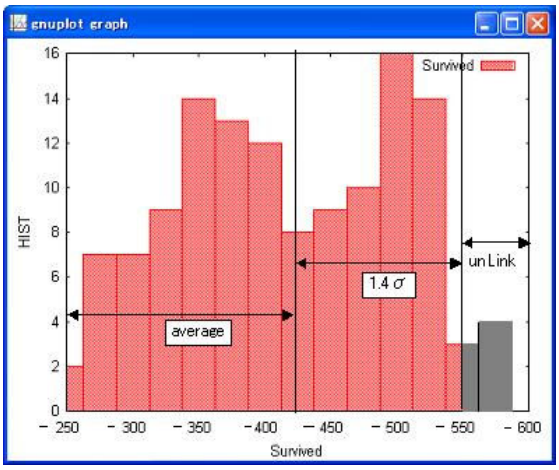
### 4.1 有意でない連結の摘出

ノードがN個ある場合、或るノードとその他のノードの連結の状態の組合わせは $2^{N-1}$ ある。例えばこの状態を(0,1,1,…0,1)と表現すると、各ノードについて連結状態の MDL を昇順で並べると下表の様になる。疎な連結が上にあり、密な連結が下になり、疎な連結は有意でないことが判る。

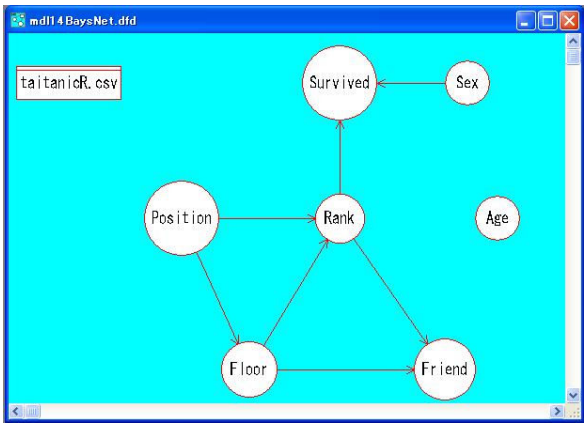
表 4-1　連結状態と MDL 値

| No | Survived | | Sex | | Rank | | Age | |
|---|---|---|---|---|---|---|---|---|
| | link | mdl | link | mdl | link | mdl | link | mdl |
| 1 | '00000000 | −593.3 | '00000000 | −578.2 | '00000000 | −888.9 | '00000000 | −1262.6 |
| 2 | '00000010 | −592.7 | '00000010 | −574.7 | '00000010 | −883.1 | '10000000 | −1254.4 |
| 3 | '00010000 | −585.1 | '00100000 | −569.7 | '01000000 | −880.4 | '01000000 | −1243.7 |
| 4 | '00010010 | −582.2 | '00000100 | −567.0 | '01000010 | −874.3 | '00000010 | −1231.5 |
| 5 | '00001000 | −550.0 | '00100010 | −566.0 | '00001000 | −845.2 | '00001000 | −1225.5 |
| 6 | '00000100 | −547.8 | '00000110 | −562.2 | '00001010 | −838.8 | '11000000 | −1222.3 |
| 7 | '00000110 | −546.9 | '00010000 | −559.3 | '10000000 | −837.1 | '00000100 | −1221.9 |
| 8 | '00000001 | −544.2 | '00000001 | −558.4 | '10000010 | −831.2 | '10000010 | −1221.0 |
| 9 | '00100000 | −541.6 | '00100100 | −558.2 | '01001000 | −830.9 | '01000010 | −1206.4 |
| 10 | '00100010 | −540.9 | '00001000 | −555.3 | '11000000 | −820.6 | '10000100 | −1202.5 |
| 122 | '01111010 | −275.2 | '10111100 | −311.6 | '01011111 | −408.4 | '11101101 | −883.6 |
| 123 | '01011111 | −269.5 | '10011111 | −298.6 | '10011111 | −404.9 | '11001111 | −870.7 |
| 124 | '01111001 | −267.0 | '10111001 | −296.9 | '11011001 | −401.1 | '11101110 | −854.2 |
| 125 | '01111110 | −256.7 | '10111110 | −287.6 | '11011101 | −396.3 | '10101111 | −843.3 |
| 126 | '01111101 | −255.7 | '10111101 | −286.8 | '11011110 | −395.8 | '01101111 | −839.0 |
| 127 | '01111011 | −239.0 | '10111011 | −273.2 | '11011011 | −370.8 | '11101011 | −815.9 |
| 128 | '01111111 | −227.9 | '10111111 | −260.6 | '11011111 | −366.8 | '11101111 | −769.1 |

例えばSurvied ノードの連結の状態のMDL 分布を見ると下図の様になっている。




ここで2点のみの連結を抽出し、90%タイル(1.4σ)以下でMDL 値が低い2点の連結（灰色の部分）は無視する方法を採る。2点のみ連結は疎なので殆どが90%タイル以下に入る。これらを非連結にすると疎だが有意な2点の連結のみ残り下図が得られる。




### 4.2 考察

ベイジアンネットを生成する学習用データ(891 件)と試験用データ(255 件)に分ける。σ の倍率を変えて Survived への連結数を増やした構造図で、生存有無の条件付確率の正解率が試験データでどう改善されているか検討してみた。

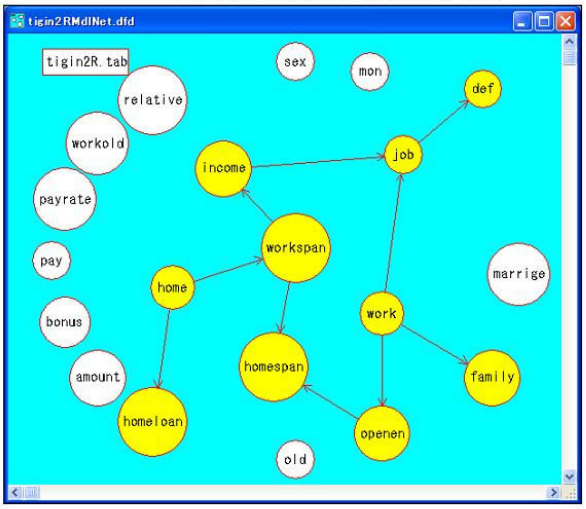
| σ | 連結数 | BIC | 正解率 |
|---|---|---|---|
| 1.0σ | 1 | 276 | 77.63% |
| 1.4σ | 2 | 240 | 78.58% |
| 1.7σ | 3 | 209 | 83.20% |
| 1.8σ | 4 | 204 | 84.82% |
| 2.0σ | 6 | 168 | 87.54% |
| ∞σ | 7 | 126 | 89.10% |

連結数が増えても生存の正解率が改善しない事が示され有意で疎なグラフでも十分データを表現していることがわかる。

## 5. まとめ

ベイジアンネットの構造推定には主に３方法あり、MDLによるスコアベースがデータを忠実に構造推定するが、組合せ爆発が発生する。これを避ける方法として、２点間のMDLが閾値以下の有意でない関係を非連結にすることにより、疎で有意なネットワークの構造推定ができることを示した。しかし実用的には閾値で判定するより、結線数を指定して最も有意な方から連結した構造図の方が理解がし易い。即ち結線数が多すぎても少なすぎても構造図は理解し難いからである。下図はローンの債務者データ6000件から結線数10本の指定で構造推定されたベイジアンネットワークである。ここでは債務不履行(def)は網掛けノードのみに関連していることがわかる。

ローン債務のベイジアンネット構造推定結果　mdl法




| | |
|---|---|
| def | 債務不履行 |
| job | 業種 |
| work | 勤務形態 |
| openen | 口座開設期間 |
| family | 家族人数 |
| homespan | 居住年数 |
| home | 住居種類 |
| homeloan | 住宅ローン有無 |
| income | 年収 |
| workspan | 勤続年数 |

構造推定された連結方向は条件付確率の対数尤度MDLが有意な場合であって、これは因果ではない。因果をデータから推定する方法として独立成分分析(ICA)を用いた方法もあるが[9]、データに無い交絡因子の存在によってデータにバイアスが発生していれば、正確な因果表現にならない。実用的には、時間軸上の生起する順番が自明なものは順番の指定で方向が順番に合う様に制御する必要がある。それでも連結方向に矛盾がある場合、隠れた因子の存在を疑う必要がある。

## 6　今後の課題

因果関係を正確に把握するには、介入操作によるデータの変化を検知する必要がある。これには時系列データの差分を利用して因果関係の強さを計測する方法が考えられる。

またベイジアンネットの構造推定をロボティクスへ応用が考えられる。これは動きのログデータから自律的に構造を構築するものである[13]。

## 謝辞

大規模ベイジアンネットワークの構造学習の研究者Dr.森下民平(きざしカンパニー)に懇切な説明を受けた。文献[8]のY.Mitsui さんのタイタニック号の構造学習の解説と公開プログラムが大変参考になった。両者にお礼を述べたい。

## 参考文献

[1] Kevin Murphy *Machine Lerning*　§20　(2012)
[2] Judea .Pearl 黒木学訳 統計的因果推論　(2009)
[3]Jie Cheng
*Learning Bayesian network form Data* (2000)
[4] 鈴木譲, 記述長最小規準と状態分割の立場から見た確率的規則の学習、電子情報通信学会論文誌A, Vol. J75-A, No. 8
[5]植野真臣 ベイジアンネットワーク (2013)
[6]Narry Wermuth,Eberhard Scheidt
*Fitting Covariance Selection Model to a Matirix*
[7]鹿島久嗣 数理情報工学特論第一
[8]Y.Mitsui データ解析の備忘録
http://mitsui725.blogspot.jp/2014/02/c.html
[9]中井眞人 独立成分分析を用いたベイジアンネット構造推定　SAS ユーザ会 (2015)
[10]C.M.Bishop
*Pattern Recognition and　Mchine Learning* §8 (2006)
[11] Daphne.Kollar,Nir Friedman
*Probabilistic Graphical Model* §3　(2009)
[12]Padoc/stat ガウシアン・モデル (2013)
http://www1.m.jcnnet.jp/mabonki/sub/ex_ggm.htm
[13]稲邑哲也 ロボティクスにおけるベイジアンネットの応用 人工知能学会誌 17 巻 5 号 (2002)