2G1-3

潜在クラスが存在する場合の ベイズ的アプローチによる非ガウス因果構造推定法

A Bayesian estimation approach for analyzing non-Gaussian data generating processes when there are latent classes

田中 直樹　清水 昌平　鷲尾 隆
Naoki Tanaka　Shohei Shimizu　Takashi Washio

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Recently, large amount of observed data has been accumulated in various fields and there is a growing need for estimating generating process of these data. It has been considered to estimate the data generating processes of variables using a linear, acyclic model based on non-Gaussianity of external influences (LiNGAM). However, results of the estimation can be biased if there are latent classes. In this paper, we first review LiNGAM, its extended model and estimation procedure for LiNGAM in a Bayesian framework. Then we propose a new Bayesian estimation procedure that solves the problem.

1. はじめに

データマイニングの分野では，データの生成モデルに非ガウス分布を仮定することで変数間の因果関係を向きも含めて導出する因果推論に関する手法が盛んに研究されている．しかし，**潜在クラス**と呼ばれる，データ生成過程の異なるデータが混在する場合，推定結果が歪められる可能性がある．近年，潜在クラスが存在する場合の観測変数間の因果関係を推定する方法が研究されているが，局所解に収束する場合がある等の問題がある [Palmer 08]．そこで本稿では非ガウス性に基づく既存手法を，潜在クラスが存在する場合でも因果構造を正しく推定できるように改良することを試みる．そして，人工的に生成したデータについて評価実験を行い，その結果について考察する．

2. 既存手法

2.1 LiNGAM モデル

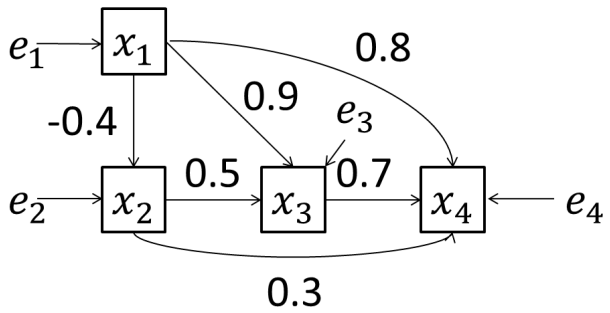

図 1: LiNGAM モデルを表す有向非巡回グラフ．ノードは変数を表し，辺は変数間の因果関係の向きとその結合の強さを表す．

まず，観測されるデータが図 1 のような有向非巡回グラフ（Directed Acyclic Graph，DAG）により生成されるものと仮定する．変数の数を q としたとき，この有向非巡回グラフを $q \times q$ の隣接行列 $\mathbf{B} = \{b_{ij}\}$ で表す．ここで b_{ij} は有向非巡回グラフにおける，変数 x_j から変数 x_i への結合の強さを表すものとする．さらに変数 x_i の因果的順序を $k(i)$ で表すこととする．加えて，変数の間の関係が線形であると仮定する．以上より，e_i を外的影響，μ_i を定数として，LiNGAM モデル [Shimizu 06] は次の式で表される．

$$x_i = \sum_{k(j)<k(i)} b_{ij}(x_j - \mu_j) + e_i + \mu_i \tag{1}$$

外的影響 e_i は全て平均 0，分散非ゼロの非ガウス分布に従う連続な確率変数である．ただし，平均や分散等の適当な次数のモーメントの存在を仮定する．また，潜在交絡変数は存在しないと仮定する．もし潜在交絡変数が存在する場合，潜在交絡変数からの影響が外的影響 e_i に含まれるため e_i は互いに独立ではない．そのため，式 (1) によって因果構造を正しく表現することができない．逆に，もし潜在交絡変数が存在しなければ e_i は互いに独立である [Spirtes 93]．

式 (1) は次の形の行列で表せる．

$$\mathbf{x} = \mathbf{B}\mathbf{x} + (\mathbf{I} - \mathbf{B})\boldsymbol{\mu} + \mathbf{e} \tag{2}$$

ここで $\mathbf{x} = [x_1, \ldots, x_q]^T$ は q 次元ベクトルであり，$\mathbf{B}$ は非巡回の仮定より，行と列を同時に並び換えることで**厳密な下三角行列**，すなわち対角要素が全て 0 の下三角行列に変形できる [Bollen 89]．

2.2 LiNGAM 混合モデル

この節では，l 個の異なる構造の LiNGAM モデルに従って生成されたデータが混在する場合を表現した LiNGAM 混合モデル [Shimizu 08] について述べる．ここで構造とは，データの変数間の結合の強さ b_{ij}，外的影響の確率密度 p_i，定数 μ_i のことを指す．LiNGAM モデルの構造が同じであるデータの集合をクラスとし，クラス c に属するデータの変数間の結合の強さを $b_{ij}^{(c)}$，外的影響を $e_i^{(c)}$，定数を $\mu_i^{(c)}$ とすると， LiNGAM 混合モデルによって，クラス c に属するデータは行列の形で以下のように表される．

$$\mathbf{x} = \mathbf{B}^{(c)}\mathbf{x} + (\mathbf{I} - \mathbf{B}^{(c)})\boldsymbol{\mu}^{(c)} + \mathbf{e}^{(c)} \tag{3}$$

連絡先: 田中 直樹，大阪大学 産業科学研究所，567-0047 大阪府茨木市美穂ヶ丘 8-1，tanaka@ar.sanken.osaka-u.ac.jp

また，$\mathbf{x}$ の要素 x_i については，

$$x_i = \sum_{k(j)<k(i)} b_{ij}^{(c)}(x_j - \mu_j^{(c)}) + \mu_i^{(c)} + e_i^{(c)} \tag{4}$$

と表される．クラスの数が 1 ならば LiNGAM 混合モデルは LiNGAM モデルと等しい．

さらに，以下の式に従って各クラスに属する観測データの確率密度が混合されるとする．

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{c=1}^{l} p(\mathbf{x}|\boldsymbol{\mu}^{(c)}, \mathbf{B}^{(c)})p(c) \tag{5}$$

ただし，$\mathbf{\Theta} = [\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(l)}], \boldsymbol{\theta}^{(c)} = [(\boldsymbol{\mu}^{(c)})^T, \mathrm{vec}(\mathbf{B}^{(c)})^T]^T$ であり，vec(・) は行列を列ごとに分解した後，一列目から順に上から並べて列ベクトルにする操作を表す．また，$p(c)$ は各確率密度の重み係数である．

2.3 BayesLiNGAM 法

本稿では 2.2 節で述べた LiNGAM 混合モデルに従ってデータが生成すると仮定して，ベイズ的アプローチにより推定を行うが，そのための準備として，この節ではデータは 2.1 節で述べた LiNGAM モデルに従って生成される，すなわち全て同じ構造（同じ b_{ij}, p_i, μ_i）の LiNGAM モデルに従うと仮定し，そのデータの因果関係を推定する際にベイズ推定を行う BayesLiNGAM 法 [Hoyer 09] について述べる．ベイズ推定は，ある複数の仮説の中からそれらの事後確率が最大である仮説を選択する推定法である．よって，BayesLiNGAM においては二変数間の因果関係の有無と向きを仮定すると，三種類の DAG ($G_1 : x_1 \quad x_2$ or $G_2 : x_1 \rightarrow x_2$ or $G_3 : x_1 \leftarrow x_2$) が存在するので，G_1, G_2, G_3 それぞれの事後確率を計算し，その値が最も大きい DAG を出力する．

ある因果構造を $G_m (m = 1, 2, 3)$，観測データセットを D（$D = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$，$N$ はサンプル数）とする．ここで，データの各サンプルが互いに独立であるという仮定より $p(\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}) = \prod_{n=1}^{N} p(\mathbf{x}^n)$ である．ベイズの定理より事後確率 $P(G_m|D)$ は次式で表される：

$$P(G_m|D) = \frac{p(D|G_m)P(G_m)}{p(D)} \tag{6}$$

よって式 (6) において，尤度 $p(D|G_m)$，事前確率 $P(G_m)$，事後確率を正規化する定数 $p(D)$ を求めれば事後確率が算出できる．

事前確率 $P(G_m)$ は事前情報を表す．もし事前情報が何もなければ，因果構造 G_m が三通りであることから $P(G_m) = \frac{1}{3}$ となる．

また，$p(D)$ は事後確率を正規化する定数であり，事前確率で評価したデータの実現確率（尤度）の期待値であるので，以下の式で求められる：

$$P(D) = \sum_{k=1}^{3} p(D|G_m)P(G_m) \tag{7}$$

最後に，尤度 $p(D|G_m)$ は，ある因果構造 G_m を仮定した時にデータが D である確率を表し，以下の式で求められる：

$$P(D|G_m) = \int p(D|\boldsymbol{\theta}, G_m)p(\boldsymbol{\theta}|G_m)d\boldsymbol{\theta} \tag{8}$$

ここで，$\boldsymbol{\theta}$ は LiNGAM モデルの式 (1) における b_{ij}, μ_i, そして外的影響 e_i の確率密度 p_i のパラメータを一つのパラメータにまとめた変数である．$p(\boldsymbol{\theta}|G_m)$ については，b_{ij} は標準ガウス分布に従うと仮定すること，そして以下のように p_i をモデリングすることで特定できる．

e_i の非ガウス性を表現するために，本稿ではガウス分布に形状（shape）パラメータを加えた一般化ガウス分布 (generalized gaussian distribution) を用いる．一般化ガウス分布は対称であり，あらゆるパラメータのガウス分布とラプラス分布，そして有界な区間の一様分布を含む．また，確率密度は尺度 (scale) パラメータ $\alpha_i (> 0)$，形状パラメータ $\beta_i (> 0)$ を用いて以下の式で表される：

$$p_i(e_i) = \frac{\beta_i \exp(-(|e_i|/\alpha_i)^{\beta_i})}{2\alpha_i \Gamma(1/\beta_i)} \tag{9}$$

$\Gamma()$ はガンマ関数であり，定義式は実部が正である複素数 z について，

$$\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1} dt \qquad (Rz > 0) \tag{10}$$

である．また，一般化ガウス分布の分散 σ_i^2 は以下の式で表される：

$$\sigma_i^2 = \frac{\alpha_i^2 \Gamma(3/\beta_i)}{\Gamma(1/\beta_i)} \tag{11}$$

さらに，$p(D|\boldsymbol{\theta}, G_m)$ については，サンプルが互いに独立であるという仮定より $p(\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}) = \prod_{n=1}^{N} p(\mathbf{x}^n)$ であるので，$p(\mathbf{x}|\boldsymbol{\theta}, G_m)$ を全てのサンプルについて掛け合わせればよい．そこで，$p(\mathbf{x}|\boldsymbol{\theta}, G_m)$ は LiNGAM モデルより導出すると，

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}, G_m) &= \prod_{i=1}^{n} p_i(e_i) \\ &= \prod_{i=1}^{n} p_i(x_i - \mu_i - \sum_{k(j)<k(i)} b_{ij}(x_j - \mu_j)) \end{aligned} \tag{12}$$

となり，$\boldsymbol{\theta}, G_m$ が与えられた時の観測データ $\mathbf{x}$ の確率密度が求められる．

2.4 ICA 混合モデル

この節では，潜在クラスが存在する場合の従来の因果構造探索法について触れる．式 (3) において，各クラスの観測変数間の影響の強さを要素にもつ行列 $\mathbf{B}^{(c)}$ を求めることで因果構造は推定されるが，LiNGAM 混合モデルのパラメータ推定方法として，LiNGAM 混合モデルを ICA(Independent Component Analysis[Hyvärinen 01]) 混合モデルに変形し，ICA 混合モデルの推定法によりパラメータを求めるというアプローチが提案されている [Shimizu 08]．ICA 混合モデルの推定法には，例えば [Palmer 08] がある．LiNGAM 混合モデルを ICA 混合モデルに変形するには，式 (3) を $\mathbf{x}$ について解き直し，

$$\mathbf{x} = \boldsymbol{\mu}^{(c)} + \mathbf{A}^{(c)}\mathbf{e}^{(c)} \tag{13}$$

とすればよい．ただし $\mathbf{A}^{(c)} = (\mathbf{I} - \mathbf{B}^{(c)})^{-1}$ である．[Palmer 08] の手法により $\boldsymbol{\mu}^{(c)}$ と $\mathbf{A}^{(c)}$ が得られるので，あとは $\mathbf{B}^{(c)}$ を計算により得ればよい．

3. 提案手法

3.1 データ生成モデル

提案手法においては，簡単のため，観測変数間の因果関係が存在することがわかっているものとする．すなわち，二種類の

DAG($G_2 : x_1 \rightarrow x_2$ or $G_3 : x_1 \leftarrow x_2$) のうち事後確率の高い方を出力するものとする．クラス c に属するデータ $\mathbf{x}$ の確率密度は，式 (4)，式 (12) より，

$$p(\mathbf{x}|\boldsymbol{\theta}^{(c)}) = \prod_{i=1}^{n} p_i^{(c)}(x_i - \mu_i^{(c)} - \sum_{k(j)<k(i)} b_{ij}^{(c)}(x_j - \mu_j^{(c)}))$$

と表され，これを式 (5) により全クラスの $\mathbf{x}$ の確率密度を混合すると，

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\theta}) &= \sum_{k=1}^{l} p(\mathbf{x}|\boldsymbol{\theta}^{(c)})p(c) \\ &= \sum_{k=1}^{l}\{\prod_{i=1}^{n} p_i^{(c)}(x_i - \mu_i^{(c)} \\ &\quad - \sum_{k(j)<k(i)} b_{ij}^{(c)}(x_j - \mu_j^{(c)}))\}p(c) \end{aligned} \quad (14)$$

となり，$\boldsymbol{\theta}^{(c)}, G_m(m = 2, 3)$ が与えられた時の観測データ $\mathbf{x}$ の確率密度を求めることができる．

3.2 推定手法

3.1 節のデータ生成モデルを用いて因果関係を直接ベイズ推定する問題点として，クラス数が多ければ多い程平均パラメータ $\mu_i^{(c)}$ の推定が難しくなるという事が挙げられる．それを解消するために，本研究では EM アルゴリズム [Bishop 06] によるクラスタリングを用いる．手順としては，まず外的影響 $e_i^{(c)}$ にガウス分布を仮定して EM アルゴリズムによりクラスタリングを行い，平均パラメータ $\mu_i^{(c)}$ のおおよその値を求める（なぜここではガウス分布を仮定するのかについては，3.3 節で述べる）．この際，さまざまなクラス数を仮定して EM アルゴリズムを実行し，ベイズ情報量基準 [Schwarz 78]（3.3 節参照）を用いることでクラス数の推定も同時に行う．そして，得られた平均パラメータ $\mu_i^{(c)}$ とクラス数を用いて今度は外的影響 $e_i^{(c)}$ に非ガウス分布を仮定してベイズ推定を行う．この時，平均パラメータ $\mu_i^{(c)}$ の事前分布の平均をクラスタリングによって得られた値とすることで，EM アルゴリズムで局所解に収束した場合でもそれを回避し，かつ EM アルゴリズムを用いず直接パラメータ推定を行う場合よりも効率的な推定が行える．

3.3 EM アルゴリズムとベイズ情報量基準

EM アルゴリズムでは，負担率（観測データが各クラスから生成された確率）を重みとして，観測データと潜在変数（データの各サンプルがどのクラスに属するのかを表す変数）の同時分布に対して最尤推定を行う．ただし，負担率とパラメータの更新を同時に行うのは困難であるので，それらを交互に更新し，収束するまで更新を行う．パラメータの更新については，観測データと潜在変数の同時分布を各パラメータについて微分し，それを 0 とおくことで最適なパラメータを求めるが，指数型分布族のうち，区間 $[-\infty, \infty]$ をサポートし，かつ連続な非ガウス分布で，各パラメータを微分してそれを 0 とおいて解くことができる分布は存在しない [Bishop 06]．よってここでは非ガウス分布ではなくガウス分布を用いる．次にベイズ情報量基準（Bayesian Information Criterion, BIC）について説明する．EM アルゴリズムにおいてはクラス数が与えられた元で計算を行い，負担率やパラメータの値と共に尤度も求められる．クラス数をデータから推定したい場合，この尤度を元に推定すればよいが，単に様々なクラス数を与えた場合の尤度のみで比較すると，クラス数が多ければ多いほどモデルの自由度が高くなり，データへの当てはまりが良く，尤度が高くなるためクラス数は多くなりがちである．そこで，モデルの自由度に対して罰則項を加えることで，できるだけ自由度の低い（簡単な）モデルの中で，データへの当てはまりの良いモデルを選ぶための評価関数としてベイズ情報量基準を用いる．そしてそれは以下のような式で表される．

$$BIC = -2\log L + a\log N \quad (15)$$

ただし，L は尤度，a はモデルの自由度，N は観測データのサンプルサイズを表す．第一項は負の対数尤度を表し，第二項はモデルの自由度に対する罰則項であるため，BIC が小さい程良いモデルであるということになる．

3.4 ディリクレ分布と多項分布

3.1 節でモデリングした，クラス毎に異なるデータの確率密度を混合する際の各確率密度の重み係数 $p(c)$ を決定するために本研究ではディリクレ分布を用いる．ディリクレ分布からサンプルを採るには，ガンマ分布に従う独立なサンプル $\gamma_1, \ldots, \gamma_Q$ を発生させて，それらの和が 1 になるように，

$$p_r = \frac{\gamma_r}{\sum_{r=1}^{Q} \gamma_r} \quad (16)$$

と正規化すればよい．ここで，ガンマ分布の確率密度関数は形状パラメータ $u > 0$，尺度パラメータ $v > 0$ の二つのパラメータを用いて次のように表される：

$$f(x) = x^{u-1}\frac{\exp(-x/v)}{\Gamma(u)v^u} \qquad (x > 0) \quad (17)$$

$\Gamma()$ はガンマ関数であり，前述の式 (10) で表される．

さらに，ディリクレ分布から得た重み係数 $p(c)$ をパラメータとして，多項分布によりデータがどのクラスに属するのかを決定する．多項分布は一般に，試行回数を N，事象 X_r が起こる確率を $p_r(r = 1, \ldots, Q)$ とすると以下の式で表される：

$$\begin{aligned} &P(X_1 = n_1, X_2 = n_2, \ldots, X_Q = n_Q) \\ &\quad = \frac{N!}{n_1!n_2!\ldots n_Q!} p_1^{n_1} p_2^{n_2} \ldots p_Q^{n_Q} \end{aligned} \quad (18)$$

ここで，p_r はディリクレ分布から得た重み係数 $p(c)$ と対応している．また，$p_r > 0, p_1 + p_2 + \cdots + p_Q = 1, n_1 + n_2 + \cdots + n_Q = N$ である．

3.5 階層ベイズ法

本稿では観測変数間の結合の強さ $b_{ij}^{(c)}$，定数 $\mu_i^{(c)}$ の事前分布としてガウス分布を用いる．その際ガウス分布のパラメータ（平均と分散）を決めなければならないが，平均は 0 とし，分散にはハイパーパラメータを用いる．ハイパーパラメータとは，分布のパラメータを規定するパラメータである．すなわち，分散を定数にするのではなく，さらに事前分布を仮定する．そして，最も事後確率を高くするパラメータを推定に利用する．このように，事前分布にさらに事前分布を仮定するようなモデルは階層ベイズ法 [伊庭 04] と呼ばれ，観測データから適切だと思われるパラメータを推定し，利用することができる．なお，本稿ではハイパーパラメータは逆ガンマ分布に従うとする．

表 1: 真のクラス数が 2 の時の実験結果．数値は 100 回中正しく推定した回数を示す．

	サンプル数		
$l = 2$	50	100	200
提案手法	87	93	94

表 2: 真のクラス数が 4 の時の実験結果．数値は 100 回中正しく推定した回数を示す．

	サンプル数		
$l = 4$	50	100	200
提案手法	84	90	96

4. 評価実験

本稿ではサンプルサイズ $N = 50, 100, 200$，潜在クラス数 $l = 2, 4$，真のグラフが $G_2 : x_1 \to x_2$ であるデータセットを，LiNGAM 混合モデル（式 (3)）に従ってそれぞれ 100 個ずつ生成した．外的影響 $e_i^{(c)}$ はそれぞれラプラス分布，一様分布，t 分布（全て平均 0，分散 1）のうちランダムでいずれかに従って生成させた．また，観測変数間の結合の強さは全クラス共通で $b_{21}^{(c)} = 0.8$ とした．さらに，外的影響の確率密度 $p_i^{(c)}$ に用いる一般化正規分布のパラメータ（式 (9) の α_i, β_i）と，$b_{ij}^{(c)}, \mu_i^{(c)}$ の事前分布（ガウス分布）の分散に，形状・尺度パラメータともに 3 である逆ガンマ分布を用いた．そしてデータセットそれぞれについて提案手法を用いて推定を行い，正しく推定できたかどうかを判定した．また，EM アルゴリズムにおいて，潜在クラスの数 $c = 1, ..., 2\log N$（小数点以下切り捨て）それぞれについてクラスタリングを行った．

実験結果を表 1 と表 2 に示す．実験結果より提案手法は，潜在クラスが存在する場合でも精度良く推定を行えることがわかった．本研究の発表当日は，従来法との比較結果も報告する．

5. 結論

本稿では，因果推論分野における既存の非ガウス性に基づく手法に，潜在クラスが存在しても正しい推定を可能にする改良を加え，その性能を評価した．従来の非ガウス性に基づく手法では，局所解に収束する等の問題点がある．本稿の成果により，潜在クラスが存在する場合でも精度良く推定を行えることを確認できた．

今後の課題としては，人工的に生成したデータではなく，現実に蓄積されたデータに対して提案手法を適用し性能評価を行うことが挙げられる．

参考文献

[Bishop 06] Bishop, C. M.: *Pattern recognition and machine learning*, Springer (2006)

[Bollen 89] Bollen, K. A.: *Structural Equations with Latent Variables*, John Wiley & Sons (1989)

[Hoyer 09] Hoyer, P. O. and Hyttinen, A.: Bayesian discovery of linear acyclic causal models, in *Proc. 25th Conf. on Uncertainty in Artificial Intelligence (UAI2009)*, pp. 240–248 (2009)

[Hyvärinen 01] Hyvärinen, A., Karhunen, J., and Oja, E.: *Independent component analysis*, Wiley, New York (2001)

[Palmer 08] Palmer, J. A., Makeig, S., Kreutz-Delgado, K., and Rao, B. D.: Newton method for the ICA mixture model, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP2008)*, pp. 1805–1808 (2008)

[Schwarz 78] Schwarz, G.: Estimating the dimension of a model, *The annals of statistics*, Vol. 6, No. 2, pp. 461–464 (1978)

[Shimizu 06] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A.: A linear non-gaussian acyclic model for causal discovery, *J. Machine Learning Research*, Vol. 7, pp. 2003–2030 (2006)

[Shimizu 08] Shimizu, S. and Hyvärinen, A.: Discovery of linear non-gaussian acyclic models in the presence of latent classes, in *Proc. 14th Int. Conf. on Neural Information Processing (ICONIP2007)*, pp. 752–761 (2008)

[Spirtes 93] Spirtes, P., Glymour, C., and Scheines, R.: *Causation, Prediction, and Search*, Springer Verlag (1993), (2nd ed. MIT Press 2000)

[伊庭 04] 伊庭 幸人, 石黒 真木夫, 松本 隆, 乾 敏郎, 田邉 國士：階層ベイズモデルとその周辺, 岩波書店 (2004)