様式 C-19

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年４月１１日現在

機関番号：１４４０１
研究種目：若手研究（B）
研究期間：2009～2011
課題番号：21700302
研究課題名（和文）　信頼性を重視した大規模変数次元小標本因果ネットワーク推定法の開発

研究課題名（英文）　Discovery of reliable causal structures in high-dimensional data

研究代表者
清水　昌平（SHIMIZU SHOHEI）
大阪大学・産業科学研究所・助教
研究者番号：10509871

研究成果の概要（和文）：大規模変数次元かつ小標本のデータから，因果ネットワークに関する知識を発見する統計解析法を開発した．具体的には，(1) 連続変数の線形因果ネットワークにおいて，因果的連鎖のトリガーの役割を果たす外生変数の推定法の開発，(2) 推定される外生変数を起点とした部分ネットワーク推定法の開発，(3) バイオインフォマティクス・ニューロインフォマティクス・社会科学などの実データによる性能評価実験を行った．また，ソフトウェアをインターネット上で公開し，成果を広く利用可能にした．

研究成果の概要（英文）：We developed several statistical methods to estimate causal networks from high-dimensional data and obtain useful causal knowledge. Specifically, we (1) developed two methods to find exogenous variables that trigger causal chains, (2) developed a direct method to estimate the entire or sub-network based on the methods for finding exogenous variables, and (3) evaluated our methods based on simulations on artificial data and real-world datasets including gene-expression data, brain imaging data and sociology data. We further made some software to perform the methods available on the internet so that many practitioners can use our methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,100,000	330,000	1,430,000
２０１０年度	900,000	270,000	1,170,000
２０１１年度	800,000	240,000	1,040,000
総　計	2,800,000	840,000	3,640,000

研究代表者の専門分野：統計的因果推論
科研費の分科・細目：統計科学・多変量解析
キーワード：統計的因果推論、構造方程式モデル、独立成分分析、因果構造探索

１．研究開始当初の背景

(1) 経緯：マイクロアレイによる遺伝子発現データでは，被験者数は数十から高々数百程度であるのに対して，遺伝子の数は少なくとも数千以上にもなる．構造方程式モデリングやベイジアンネットワーク等の従

来の統計的因果分析(Bollen, Wiley, 1989; Pearl, Cambridge Univ. Press, 2000)は，標本の大きさが変数次元より十分大きいことを前提としている．そのため，単にそのまま大規模変数次元データに適用しても，推定結果を出力することさえできないことも多い．標本の大きさが小さすぎて，推定に用いることのできる情報が少なすぎるからだ．そこで，現在の研究の流れでは，背景情報等を用いて，足りない分の情報を補おうとする(井元，東京電機大学出版，2007，第 4 章)．しかし，情報の不足を十分に補えることはあまりなく，推定結果の信頼性が必ずしも高くないという問題を抱えている．大規模変数次元かつ小標本のデータが日々蓄積されているにもかかわらず，このようなデータから因果構造に関する信頼できる知識を抽出できる統計解析法はまだ出来上がっていない．しらみつぶしに実験を繰り返すことはコストや時間の面からも非常に難しい．効率の良い実験を行うためにも，非実験データから因果ネットワーク構造に関する有望な仮説(知識)をデータから引き出す統計解析法が求められている．同様の問題は，ニューロインフォマティクスにおいて，機能的磁気共鳴画像(fMRI)データ等によって脳の各部位の結合性を解析するような場合にも起こっている(Londei et al., Cognitive processing, 2006)．また，社会科学の質問紙調査における構成概念間の因果分析などでも，標本の大きさが十分大きいかについて議論になることも多く(Bollen, Wiley, 1989)，標本の大きさがあまり大きくない状況でも信頼できる統計解析法の需要は大きい．

(2) 動機：構造方程式モデリングやベイジアンネットワーク等の従来の(連続変数の)統計的因果分析法は，陰に陽に正規性の仮定に基づいている．そのため，データが非正規分布に従う場合であっても，共分散行列の持つ情報しかモデル識別に用いて来なかった．一方，信号処理の分野で提案された独立成分分析(Hyvarinen et al., Wiley, 2001)は，観測変数に非正規性が認められる場合に，その非正規性をモデル識別に活用する．申請者は，この「データの非正規性の活用」という独立成分分析のアイデアを統計的因果推論に持ち込み，それまで識別不可能であった因果モデルの多くを識別できる統計解析法を開発してきた．例えば，最も単純な 2 変数の例を挙げる．モデル 1(x2 := b21 x1 + e2, b21 は影響の強さを表す定数，e2 は誤差変数) が真のデータの発生機構を表しているとしよう．モデル 1 と，モデル 2 (x1:= b12 x2 + e1, b12 は影響の強さを表す定数，e1 は誤差変数) とは生成順序が反対である．従来の共分散行列に基づく方法では，(非実験)データからモデル 1 とモデル 2 を区別することはできない(Bollen, Wiley, 1989)．どちらも同等にデータ共分散行列に適合するからだ．しかし，データの非正規性，つまり，共分散行列以外の情報も用いることで，区別できることがわかっている．このように，データの非正規性を活用することで，共分散行列に基づく従来法では不可能だった分析が可能になることがわかってきた．また，バイオインフォマティクスやニューロインフォマティクスにおいて，上述のような変数次元と標本の大きさのアンバランスさによる深刻な問題があることを，応用研究者と議論する中で知った．そこで，非正規性の活用による高いモデル識別能力を武器にして，「変数次元が標本の大きさより遥かに大きい」という大規模変数次元データ解析における統計的課題の解決に貢献したいと考えるに至った．

２．研究の目的

大規模変数次元かつ小標本のデータから，因果ネットワークに関する信頼性の高い知識を発見する統計解析法を開発する．ネットワークのモデルとしては，連続変数の線形モデルを基本とする．線形モデルは，例えば構造方程式モデリングの分野において多くの先行研究があり，応用分野で成果を挙げている(Bollen, Wilely, 1989; Kim et al., Human Brain Mapping, 2007)．そこで，線形モデルに関する結果を基礎にして，実データに適用する中で，応用分野に適した方向にモデルを拡張していくことにする．

具体的には，(1) 連続変数の線形因果ネットワークにおいて，因果的連鎖のトリガーの役割を果たす外生変数の推定法の開発，(2) 推定される外生変数を起点とした部分ネットワーク推定法の開発，(3) 現実の問題への応用(バイオインフォマティクス・ニューロインフォマティクス・社会科学)の３段階の研究を行う．また，ソフトウェアをインターネット上で適宜公開する．

３．研究の方法

次の 3 段階に分けて、研究を行った．

(1) 外生変数の推定法の開発：小標本でも信頼性高く推定できるような小さくかつ重要な部分として，まず因果ネットワークの外生変数に着目する．外生変数は，因果システムのトリガーの役割を果たすと考えられる．外生変数を同定することにより，効果的にシステムを制御したり，現象の根本原因の同定に関する仮説を得たりできると予想される．例えば，遺伝子レベルでのダイオキシンのマウスの肝臓への影響を評価したいとしよう．遺伝子ネットワークには多

くの遺伝子が含まれているが，これら遺伝子の中で，最初にダイオキシンが影響を与える遺伝子が，外生変数にあたる．ダイオキシンの影響は，この外生変数にあたる遺伝子を経由し，ネットワーク上の他の遺伝子に伝播していく．したがって，それら外生変数にあたる遺伝子を守ることができれば，他の遺伝子への影響を防ぐことができるかもしれない．このような外生変数の推定を，第１段階の目標とする．

(2) 推定された外生変数を起点とした部分ネットワーク推定法の開発: 第 2 段階としては，推定された外生変数を起点として，部分ネットワークを推定する統計解析法を研究開発する．いろいろな方向があるが，アイデアの１つを簡単に述べる．まず，推定された外生変数の影響をデータから取り除く(回帰分析などを用いて可能)．そうして出来た「新たな」線形因果ネットワークに，(1)で開発する外生変数の探索法をもう一度適用する．これを繰り返し，ネットワークの上流から探索範囲を順に広げていく方向が考えられる．

(3) 実データによる妥当性検証と推定法の拡張: 次年度以降は，初年度に開発する線形因果ネットワークの探索法の仮定を緩め，各応用分野に適した方向に拡張し，より柔軟に因果関係をモデリングできるようにする．例えば，未観測交絡変数の存在の可能性・線形モデルでは不十分・異質な母集団の混合等の問題が考えられる．研究協力者からバイオインフォマティクス，ニューロインフォマティクス，社会科学の実データの提供を受け，優先度の高い問題点を洗い出し，効率的に研究を進める．従来の設定(変数次元より標本の大きさが十分大きい)における非正規性を利用する因果分析法において，これらの仮定のくずれへの対処法が今盛んに研究されている．これら最新の結果を参考にするとともに，新しいアイデアを探索し，少しでも現実に即した柔軟なモデリングができるようにしていく．

４．研究成果

(1) 外生変数の同定法

大規模変数次元小標本データにおいて外生変数を推定する方法を開発した．思い切ってモデルを簡略化することによって，計算時間を大幅に短縮し，また推定結果を安定させることに成功した．その代り，モデルの一般性は幾分低下したが，数値実験や実データによる検証実験においては致命的なものにはならなかった．また、大規模変数次元小標本データに特化しているわけではないが，より一般的な設定(LiNGAM モデル)で外生変数を同定する方法も開発した．

(2)部分ネットワーク推定法

(1)で開発した方法を基に、LiNGAM モデルにおける部分ネットワーク推定法を開発した．既存の方法の多くは，初期値の設定やステップサイズ，収束基準の選定が必要になる．しかし，それらを適切に設定することが難しい状況がよくある．本研究で開発した推定法は，その類のチューニングが必要ない直接法である．数値実験及び実データによる検証実験を行い，既存手法よりも優れたパフォーマンスを確認した．

(3) 時間構造の利用 ：

LiNGAM は，時間構造のないデータを対象としているが，時間構造のあるデータも多く存在する．時間構造のある場合の分析法の定番には，自己回帰モデル(AR モデル)がある．この AR モデルと LiNGAM モデルを組み合わせて，AR-LiNGAM を開発し，識別性の証明と推定法の提案を行った．この AR-LiNGAM は Econometrics においてよく知られている Structural Vector Auto-Regression model の非ガウス版と考えれる．また，AR モデルを自己回帰移動平均モデル(ARMA モデル)に置き換え，時間方向に潜在変数を許すモデルも開発した．

(4) 複数の LiNGAM モデルの同時推定:

LiNGAM モデルを拡張し，多母集団の同時分析を行うためのフレームワークを開発した．複数の実験条件取得された遺伝子発現データを効果的に融合して推定精度高めたり，集団に共通する因果構造や，集団ごとに異なる因果構造を推定したりできるようになった．また，同様のことが，複数被験者から取得された脳活動計測データから脳領域ネットワークを推定するためにも行える．

(5) LiNGAM の結果の信頼性評価:

LiNGAM の推定結果のばらつきを評価する方法を開発した．これにより，推定結果の統計的信頼性を吟味することができるようになった．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

① S. Shimizu. Joint estimation of linear non-Gaussian acyclic models. Neurocomputing, 査読有, 81: 104-107, 2012,DOI:10.1016/j.neucom.2011.11.005

② S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P.

O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 査読有, 12(Apr): 1225-1248, 2011,URL:http://jmlr.csail.mit.edu/papers/volume12/shimizu11a/shimizu11a.pdf

③ Y. Sogawa, S. Shimizu, T. Shimamura, A. Hyvarinen, T. Washio, and S. Imoto. Estimating exogenous variables in data with more variables than observations. Neural Networks, 査読有, 24(8): 875-880,2011,DOI:10.1016/j.neunet.2011.05.017

④ Y. Kawahara, S. Shimizu, and T. Washio. Analyzing relationships among ARMA processes based on non-Gaussianity of external influences. Neurocomputing, 査読有, 74(12-13): 2212-2221, 2011, DOI:http://dx.doi.org/10.1016/j.neucom.2011.02.008

⑤ A. Hyvarinen, K. Zhang, S. Shimizu and P. O. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. Journal of Machine Learning Research, 査読有, 11(May): 1709-1731, 2010, URL:http://jmlr.csail.mit.edu/papers/volume11/hyvarinen10a/hyvarinen10a.pdf

〔学会発表〕(計２０件)

① 清水昌平、複数データセットによる非ガウス構造方程式モデルの推定、情報統計力学の最前線ー情報と揺らぎの制御の物理学を目指してー、2012年3月21日、京都大学（京都） （招待講演）

② 清水昌平、複数データセットによる非ガウス構造方程式モデルの推定、科研費シンポジウム「生体数理・社会数理の統計科学」、2012年3月2日、早稲田大学（東京）

③ 鈴木譲, 清水昌平, 鷲尾隆、離散データの因果の同定 ～ 2値から、多値への一般化について ～、第14回情報論的学習理論ワークショップ、2011年11月10日、奈良女子大学（奈良）

④ 稲積孝紀, 鷲尾隆, 清水昌平, 鈴木譲, 山本章博, 河原吉伸、分割表の独立性に基づく二値データ生成過程の推定法、第14回情報論的学習理論ワークショップ、2011年11月10日、奈良女子大学（奈良）

⑤ T. Inazumi, T. Washio, S. Shimizu, J. Suzuki, A. Yamamoto and Y. Kawahara. Discovering causal structures in binary exclusive-or skew acyclic models. 27th Conf. on Uncertainty in Artificial Intelligence (UAI2011), 2011年7月16日, バルセロナ(スペイン)

⑥ Marina Domeshko, 鷲尾隆, 河原吉伸, 清水昌平、Analyzing relationships between CTARMA and ARMA models、第25回人工知能学会全国大会、2011年6月2日、いわて県民情報交流センター（盛岡）

⑦ 田代竜也, 清水昌平, 河原吉伸, 鷲尾隆、定常時系列データの非ガウス性を用いたARMAモデルによる変数間決定関係の解析、第25回人工知能学会全国大会、2011年6月2日、いわて県民情報交流センター（盛岡）

⑧ 稲積孝紀, 鷲尾隆, 清水昌平, 鈴木譲, 山本章博, 河原吉伸、二値データに対するデータ生成過程の推定、第25回人工知能学会全国大会、2011年6月2日、いわて県民情報交流センター（盛岡）

⑨ 清水昌平、構造方程式モデルによるデータ生成過程の学習, 特に非ガウス性の利用、第13回情報論的学習理論ワークショップ(IBIS2010)、2010年11月4日、東京大学（東京）

⑩ T. Inazumi, S. Shimizu and T. Washio. Use of prior knowledge in a non-Gaussian method for learning linear structural equation models. 9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA2010),

2010年月9月28日，サンマロ(フランス)

⑪ Y. Sogawa, S. Shimizu, A. Hyvarinen, T. Washio, T. Shimamura and S. Imoto. Discovery of exogenous variables in data with more variables than observations. 20th Int. Conf. on Artificial Neural Networks (ICANN2010), 2010年月9月17日，テッサロニキ(ギリシャ)

⑫ Y. Komatsu, S. Shimizu and H. Shimodaira. Assessing statistical reliability of LiNGAM via multiscale bootstrap. 20th Int. Conf. on Artificial Neural Networks (ICANN2010), 2010年9月15日，テッサロニキ(ギリシャ)

⑬ 小松勇介、下平英寿、清水昌平、ブートストラップ確率の計算誤差を修正するためのマルチスケール・ブートストラップ法： LiNGAM因果構造推定の場合、2010年度 統計関連学会連合大会、2010年9月6日、早稲田大学（東京）

⑭ Y. Sogawa, S. Shimizu, Y. Kawahara and T. Washio. An experimental comparison of linear non-Gaussian causal discovery methods and their variants. Int. Joint Conf. on Neural Networks (IJCNN2010), part of the IEEE World Congress on Computational Intelligence (WCCI2010), 2010年月7月23日，バルセロナ(スペイン)

⑮ S. Shimizu and Y. Kawahara. Non-Gaussian methods for learning linear structural equation models. 26th Conference on Uncertainty in Artificial Intelligence (UAI2010), 2010年月7月8日，カタナリーナ島(米国) (招待チュートリアル)

⑯ 稲積孝紀、十河泰弘、清水昌平、河原吉伸、鷲尾 隆、データの非正規性を活用する因果構造探索法と事前情報の利用、第24回人工知能学会全国大会、2010年6月9日、長崎ブリックホール（長崎）

⑰ 小松勇介、清水昌平、下平英寿、マルチスケール・ブートストラップを用いた信頼度計算： LiNGAMによる因果モデル探索の場合、2009年度 統計関連学会連合大会、2009年9月7日、同志社大学（京都）

⑱ S. Shimizu, A. Hyvärinen, Y. Kawahara and T. Washio. Identification of an exogenous variable in a linear non-Gaussian structural equation model. The Fourth International Workshop on Data-Mining and Statistical Science (DMSS2009), 2009年7月8日，京大会館(京都)

⑲ S. Shimizu, A. Hyvarinen, Y. Kawahara and T. Washio. A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model. 25th Conf. on Uncertainty in Artificial Intelligence (UAI2009), 2009年6月21日，モントリオール(カナダ)

⑳ 十河泰弘、清水昌平、鷲尾 隆、井元清哉、独立成分分析を用いた外生的発現遺伝子同定解析、第23回人工知能学会全国大会、2009年6月18日、サンポート高松（高松）

〔その他〕
ホームページ等
http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/

ソフトウェア
DirectLiNGAM:
http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/code/Dlingamcode.html

JointDirectLiNGAM:
http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/code/jointDlingamcode.html

ARMA-DirectLiNGAM:
http://www.ar.sanken.osaka-u.ac.jp/~sshimizu/code/armaDlingamcode.html

Bexsam:
http://www.ar.sanken.osaka-u.ac.jp/~inazumi/bexsam.html

６．研究組織

(1)研究代表者

清水　昌平（SHIMIZU　SHOHEI）

大阪大学・産業科学研究所・助教

研究者番号：10509871

(2)研究分担者

該当なし

(3)連携研究者

該当なし