DAGs with NO TEARS：
構造学習のための連続的な最適化

Xun Zheng[1]**、Bryon** Aragam[1]**、Pradeep** Ravikumar[1]**、Eric P.** Xing[1,2]
[1]カーネギーメロン大学　　[2]株式会社ペチューム
{xunzheng,naragam,pradeepr,epxing}@cs.cmu.edu

アブストラクト

有向非循環グラフ（DAG、ベイジアン ネットワークとしても知られる）の構造を推定することは、DAGの探索空間が組合せ的であり、ノード数に応じて超指数関数的にスケールするため、困難な問題である。既存のアプローチは、非周期性制約を強制するための様々な局所的ヒューリスティックに依存している。本論文では、この組み合わせ制約を完全に回避するために、構造学習問題を実数行列に対する純粋な*連続*最適化問題として定式化することで、根本的に異なる戦略を導入する。これは、滑らかであるばかりでなく、厳密でもある非周期性の新しい特徴付けによって達成される。その結果、この問題は標準的な数値アルゴリズムで効率的に解くことができ、実装も容易となる。提案手法は、グラフに構造的な仮定（樹幅や次数の制限など）を課すことなく、既存の手法を凌駕する。

1　はじめに

有向無尽グラフ（DAG）をデータから学習することはNP困難な問題である[8, 11]が、これは主に、効率的に実施することが困難な組合せ無サイクル性制約に起因している。一方で、DAGは生物学[33]、遺伝学[49]、機械学習[22]、因果推論[42]などに応用されており、実際によく使われるモデルである。このため、DAGを学習するための新しい手法の開発は、機械学習や統計学の中心的な課題として残されている。

本論文では、従来の*組合せ*最適化問題（左）を*連続*プログラム（右）に変換することで、DAGのスコアベース学習のための新しいアプローチを提案します：

$$\min_{W \in \mathbb{R}^{d\times d}} F(W) \quad \Leftrightarrow \quad \min_{W \in \mathbb{R}^{d\times d}} F(W) \tag{1}$$

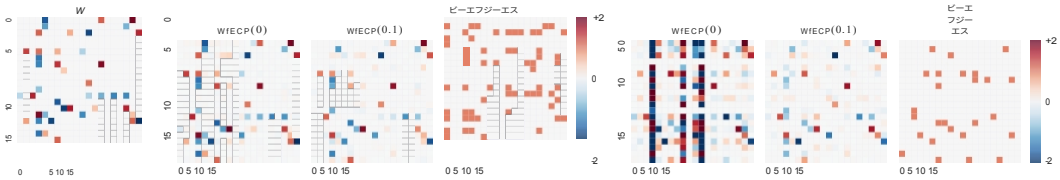
$G(W) \in$ DAGsを対象とする。　　$h(W)=0$を条件とする、

ここで、G(*W*)は重み付き隣接行列*W*が誘導するd-*node*グラフ、$F : \mathbb{R}^{d \to d}$ ！ Rはスコア関数（詳細は2.1節参照）であり，我々の重要な技術的工夫 $h : \mathbb{R}^{d \to d}$ ！ Rは実数行列上の平滑関数であり、そのレベルセットが0であることが非周期グラフを正確に特徴付ける。とはいえ

2つの問題が等価である場合、右の連続プログラムは、DAGの組合せ空間の探索に合わせた特別なアルゴリズムを必要としない。その代わりに、制約付き問題のための標準的な数値アルゴリズムを活用することができ、グラフィカルモデルに関する知識を必要とせず、特に簡単に実装することができる。これは、無向グラフモデルの状況に似ており、連続ログデットプログラムの定式化である

[4]は、無向グラフの構造学習における一連の顕著な進歩に火をつけた（セクション2.2）。しかし、凸プログラムに還元できる無向モデルとは異なり、プログラム(1)は*非凸*である。しかし、このプログラムに対するナイーブな解法でも、DAGの学習において最先端の結果を得ることができることを紹介する。

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.



(a) 真のグラフ　(b) エヌエックス　(c) $n = 20$で推定

図1：20ノードグラフの学習済み重み付き隣接行列のビジュアル比較

$n = 1000$（大きなサンプル）、$n = 20$（不十分なサンプル）：$W_{\text{ECP}}(\lambda)$は提案されたNOTEARSアルゴリズムでl_1-*正則化*λ、B_{FGS} はベースライン[31]の二値推定値である。提案されたアルゴリズムは、大きなサンプルでは良好な性能を発揮し、小さなnではl_1 正則化で精度を維持する。

貢献度本研究の主な目的は、L-BFGS [28]のような標準的な滑らかな最適化スキームを活用できるように、DAGのスコアベース学習を再形成することである。これを達成するために、我々は以下の具体的な貢献を行う：

- 非周期性制約を符号化する、計算可能な導関数を持つ$\mathbb{R}^{d\to d}$ 上の滑らかな関数を明示的に構築する。これにより、(4)の組合せ制約G 2 Dを滑らかな等式制約に置き換えることができる。
- 高次元のデータから疎なDAGの構造とパラメータを同時に推定するための等式制約付きプログラムを開発し、標準的な
  数値ソルバーで定常点を求めることができる。
- 我々は、既存の最先端技術に対する実証評価で、得られた方法の有効性を実証する。簡単な図解は図1、詳細はセクション5を参照。
- 私たちは、私たちの成果を厳密なグローバルミニマイザー[12]と比較し、私たちの方法が、実際にグローバルに最適なスコアに匹敵するスコアを達成することを示しました。
  は、定常点を見つけることだけが保証されています。

最も興味深いのは、我々のアプローチが非常にシンプルであり、約50行のPythonコードで実装できることである。そのシンプルさと実装の容易さから、我々はこの手法をNOTEARS: *Non Combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning*と呼んでいます。実装は `https://github.com/xunzheng/ notears` で公開されています。

2　背景

基本的なDAG学習問題は、以下のように定式化される：$\mathbb{R}^{n\to d}$ を、ランダムベクトル$X = (X_1, \ldots, X_d)$ の n *個の* i.i.d. 観察からなるデータ行列とし、D を、d *個のノード上の* DAG G = (V, E) の（離散）空間とする。**が**与えられたとき、我々は共同分布P （X）に対するDAG G 2 D（*ベイジアンネットワークとも呼ばれる*）を学習することを求める[22, 42]。X *を*構造方程式モデル(SEM)でモデル化する。

加重隣接行列 $W \in \mathbb{R}^{d \times d}$ で定義される。このように、離散空間上で操作するのではなく $D, \mathbb{R}^{d \times d}, d \to \infty$ 実数行列の連続空間上で操作することにする。

2.1 スコア関数とSEM

任意の $W \in \mathbb{R}^{d \times d}$ は、次のように d ノード上のグラフを定義する：$A(W) \in \{0, 1\}^{d \times d}$ を $[A(W)]_{ij} = 1 \iff w_{ij} \neq 0$、それ以外は0となるような二値行列とすると、$A(W)$ は有向グラフ $G(W)$ の隣接行列を定義している．表記を少し乱用して、W を（加重）グラフであるかのように扱うことにする。グラフ $G(W)$ に加えて、$W = [w_1 | \cdots | w_d]$ は、$X_j = w_j^T X + z_j$ によって線形SEMを定義し、$X = (X_1, \ldots, X_d)$ はランダムベクトル、$z = (z_1, \ldots, z_d)$ はランダムノイズベクトルである。z がガウス型であることは仮定しない。より一般的には、X_j のモデル化には一般化線形モデル（GLM）$\mathbb{E}(X_j \mid X_{pa(X_j)}) = f_j(w_j^T X)$。例えば、$X_j \in \{0, 1\}$ の場合、以下のようになります。

$$\mathbb{E}(X_j \mid X_{pa(X_j)}) = f_j(w_j^T X)$$

は、ロジスティック回帰により、親が与えられた X_j の条件付分布をモデル化する。

本論文では、線形SEMと最小二乗（LS）損失 $\ell(W; X) = \frac{1}{2n} \|X - XW\|_F^2$ に焦点を当てる、ただし，この後，$\mathbb{R}^{d \times d}$ 上で定義された任意の滑らかな損失関数 ℓ にすべて適用されます．

スコアリングDAGにおけるLS損失の統計的性質は広く研究されている：LS損失の最小化は、有限サンプルと高次元（d n）において高い確率で真のDAGを回復することが証明されており、したがって、ガウス型SEM [3, 45] と非ガウス型SEM [24] の両方において整合性がある。[1]また、これらの結果は、忠実性の仮定がこのセットアップでは必要ないことを意味していることに注意してください。統計的な問題についてのこのような広範な先行研究を考慮し、本論文では、LS損失を最小化するSEMを見つけるという計算上の問題に完全に焦点を当てる。

グラフとSEMの間のこの変換は、我々のアプローチの中心である。我々は*疎な*DAGを学習することに興味があるので、l_1 -正則化 $\|W\|_1 = \|\mathrm{vec}(W)\|_1$ その結果、正則化スコア関数を追加する。

$$F(W) = l(W; X) + \lambda \|W\|_1 = \frac{1}{2n} \|X - XW\|_F^2 + \lambda \|W\|_1. \tag{2}$$

したがって、私たちは解決策を模索します。

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) \quad \text{subject to} \quad G(W) \in \mathrm{DAG}. \tag{3}$$

残念ながら、$F(W)$ は連続的であるが、DAG 制約 $G(W) \in \mathrm{DAG}$ を実施することは困難である。第3節では、この離散的な制約を滑らかな等式
の制約を受けます。

2.2 前作

従来、スコアベース学習は、DAGの集合に対する*離散的なスコア*$Q : \mathrm{DAG} \to \mathbb{R}$を最適化することを目的としていた！これは、我々のスコア$F(W)$ のドメインがDではなく $\mathbb{R}^{d \times d}$ であることとは異なる：

$$\min_{G \in \mathrm{DAG}} Q(G) \tag{4}$$

G ∈ DAG に従う。

一般的なスコア関数には BDe(u) [20], BGe [23], BIC [10], MDL [6] がある．残念ながら，(4)は，最適化問題の非凸，組合せ的な性質から，NP-hard [8, 11]である．これが、(4)を解くための既存のアプローチの主な欠点である：非周期性制約は、非周期的な構造の数がd*に対して*超指数的に増加する組み合わせ的な制約である[32]。それにもかかわらず、小さな問題に対して(4)を大域最適に解くアルゴリズムが存在する[12, 13, 29, 39, 40, 47]。また、順序探索 [30, 34-36, 43]、貪欲探索 [9, 20, 31]、座標降下 [2, 16, 18]に基づく近似アルゴリズムについても幅広い文献が存在する。前者の順序に基づく方法は、トポロジカルな順序の空間を検索することで、非周期性を強制する困難な問題をd！順序の検索とトレードオフしているのに対し、後者の方法は、一度に1辺ずつ非周期性を強制し、辺が追加されるたびに非周期性違反を明示的にチェックする。(4)の直接最適化を避ける他のアプロ

ーチとしては、制約に基づく方法[41, 42]、ハイブリッド法[17, 44]、ベイズ法[14, 27, 51]が挙げられる。

この問題を、類似のよく知られた問題である「データから無向グラフ（マルコフネットワーク）を学習する」と比較することは有益である。無向グラフの学習には、(4)と同様の離散スコアに基づくスコアベース手法が初期に普及した[例えば22、§20.7]。さらに最近では、この問題を実数対称行列上の凸プログラムとして再定式化することにより[4, 48]、無向グラフを学習するための極めて効率的なアルゴリズムが開発されている[15, 21, 37]。この成功の重要な要因の一つは、広範な最適化文献の既存の技術を適用できる、閉形式の扱いやすいプログラムを持つことであった。しかし、DAG学習の一般的な問題では、このような恩恵は受けていない。

(4).本研究の主な目標の一つは、同様の閉形式の連続プログラムによってスコアベース学習を定式化することである。これを達成するための重要な工夫は、次節で紹介する非周期性の滑らかな特徴付けである。

---

[1]非凸のため、複数の最小化器が存在する可能性がある：これらの問題や、パラメータの同定可能性などの技術的な問題については、引用文献で詳しく述べられています。

3　非周期性の新しい性格付け

(3)をブラックボックス最適化するために，(3)の組合せ非循環性制約 G(*W*) 2 D を単一の滑らかな等式制約 *h*(*W*) = 0 で置き換えることを提案する．理想的には，関数 *h* : $R^{d\to d}$ ！R は、次のような望ましい条件を満たすものである：

(a) *h*(*W*) = 0 *W*が非周期的である（すなわちG(*W*) 2 D）場合に限り、0とする；
(b) *h*の値は、グラフの「DAGらしさ」を定量化するものです；
(c) hは滑らかである；
(d) hとその導関数は簡単に計算できる。

(b)の特性は、診断のために実際に役立つものである。DAGらしさ」とは、*W*がDから離れるにつれて非周期性の違反がどの程度深刻になるかを定量化することを意味する。Dへの「距離」の概念を測定することで(b)を満たす方法は数多くあるが、典型的なアプローチでは
(c)と(d)である．例えば、*h*はDへの最小l_2 距離であったり、*W*のすべての環状パスに沿ったエッジ重みの和であったりするが、これらは非平滑（（c）に反する）か計算が困難（（d）に反する）である。(a)〜(d)を満たす関数が存在すれば、ラグランジュ乗数など既存の制約付き最適化の仕組みを応用することが期待できる。その結果、DAG学習問題は、グラフ構造に依存しない数値最適化問題を解くことと同等になる。

我々の主な結果は、そのような関数の存在を立証するものである：

Theorem 1. *A matrix W* 2 $R^{d\to d}$ *is a DAG if and only if*

$$h(W) = \operatorname{tr} e^{W\circ W} - d = 0、 \tag{5}$$

*ここで、◦はハダマード積、e^A はAの行列指数である。また、*h*(W)は単純な勾配を持ち*

$$\nabla h(W) = \left(e^{W\circ W}\right)^{T} \circ 2W, \tag{6}$$

であり、（a）〜（d）のすべての望みを満たす。

ここでは、最初の主張の証明をスケッチする。定理1の正式な証明は、付録Aにある。*S* = *W* ◦ *W* とすると、*S* 2 $R^{d\to d}$ は、*W* のスパースパターンを保存している。任意の正の整数*k*について、行列power $(S)^k_{ij}$ のエントリは、ノードiからノード*j*へのすべての*k*ステップのパスに沿った重み積の合計であることを思い出してください。Sは非負なので、グラフにk*サイクル*が存在しない場合、 tr(S^k) = 0です。展開するパワーシリーズ

$$\operatorname{tr}(e^S) = \operatorname{tr}(I) + \operatorname{tr}(S) + \frac{1}{2!}\operatorname{tr}(S^2) + \cdots \geqq d、 \tag{7}$$

であり、*S*の基底グラフ（等価的に*W*）がサイクルを持たない場合に、等式が成立する。

これは数値解析においてよく研究されている関数であり、その $O(d^3)$ アルゴリズム [1] は多くの科学計算ライブラリで容易に利用できる。行列の累乗とグラフのサイクル数との関係はよく知られているが[19]、我々の知る限り、この非周期性の特性はDAG学習の文献に以前は現れていない。他の可能性のある特徴付けについては、付録の議論に譲る。

次節では、定理1を適用して、プログラム(3)を等式制約付きプログラムとして扱い、定常性まで解く。

4　最適化

定理1は、計算可能な非周期性の滑らかで代数的な特徴付けを確立している。その結果、以下の等式制約付きプログラム(ECP)は(3)と等価である：

(ECP)である。　　ミニ　F（W）　(8)

W 2Rd□d

h(W)=0を条件とする。

アルゴリズム1 NOTEARSアルゴリズム

1. 入力：初期推測 (W_0, ρ_0), 進捗率 $c \in (0, 1)$, 許容度$\epsilon > 0$, 閾値$\omega > 0$.
2. $t = 0, 1, 2, \ldots$ の場合：
   (a) プライマル$W_{t+1} \leftarrow \arg\min_W L^{\rho}(W, \rho_t)$ を $h(W_{t+1}) < ch(W_t)$ となるように□で解けよ。
   (b) $_t$デュアルアセント $\rho_{t+1} \leftarrow \rho h(W)$ です。$_{t+1}$
   (c) $h(W_{t+1}) < \epsilon$の場合、$W_{ECP} = W_{t+1}$ と設定し、ブレークする。
3. 閾値行列W_C := $_{WfECP}$ ◦ 1(|WfECP| $> \omega$)を返す。

(3)と(4)に比べて(ECP)の主な利点は、数理最適化の文献にある古典的な手法に従順であることです。しかし、$\{W : h(W) = 0\}$は非凸制約であるため、(8)は非凸プログラムであり、非凸プログラムの難しさを引き継いでいる。
の最適化を行う。特に、(8)の定常点を見つけることに満足する。5.3節では、我々の結果を大域最小化と比較し、我々の方法で見つけた定常点が、実際には大域最小化に近いことを示す。

以下では、(8)を解くためのアルゴリズムを概説する。このアルゴリズムは3つのステップから構成されている：(i) *制約付き*問題を一連の*非制約付き*部分問題に変換する、(ii) 非制約付き部分問題を最適化する、(iii) 閾値を設ける。完全なアルゴリズムはアルゴリズム1に概説されている。

4.1 拡張ラグランジアンによるECPの解法

(ECP)の解法には、二次関数のペナルティで補強された元の問題を解く拡張ラグランジュ法[例えば25]を使用する予定です：

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) + \frac{\rho}{2} |h(W)|^2 \quad \text{h}(W) = 0\text{に従う} \tag{9}$$

拡張ラグランジュ法の優れた特性は、ペナルティ*パラメータ*ρを無限大にすることなく、*制約付き*問題の解を*無制約*問題の解でうまく近似することです[25]。このアルゴリズムは、基本的に(9)の双対上昇法である。まず、ラグランジュ*乗数*αを持つ双対関数は次式で与えられる。

$$D(\alpha) = \min_{W \in \mathbb{R}^{d \times d}} L^{\rho}(W, \alpha) \text{です、} \tag{10}$$

$$\text{ここで、} L^{\rho}(W, \alpha) = F(W) + \frac{\rho}{2} |h(W)|^2 + \alpha h(W) \tag{11}$$

は補強されたラグランジアンである。目標は、二重問題の局所解を見つけることである。

$$\max_{\alpha \in \mathbb{R}} D(\alpha). \tag{12}$$

$W^?$ を□でのラグランジアン(10)の局所最小化器、すなわち $D(□)= L^{→}(W^?, □)$ とする。となるので、デュアル
目的$D(□)$は□に線形であり、微分は単に$rD(□)=h(W^?)$ で与えられる。したがって、次のようになります。
は、(12)を最適化するためにデュアルグラジエントアセントを実行する：

$$□ ← □h(W^?)\text{となります、} \tag{13}$$

ここで、ステップ*サイズ*□の選択には、以下の収束率が付属しています：

命題1(Corollary 11.2.1, 25)**。が***十分に大きく、解の近くの始点*□$_0$ *の場合。*
$^?□$ 、*更新*（13）*は線形に*$□^?$ *に収束する。*

私たちの実験では、通常、拡張ラグランジュスキームの10ステップ未満しか必要とされません。

4.2 無制約サブ問題の解決

拡張ラグランジアンは，*制約付き*問題(9)を一連の*無制約問題(10)*に変換する．
の問題(10)を解くことができる。次に、これらの部分問題を効率よく解く方法について説明する。$\boldsymbol{w} = \mathrm{vec}(W) \in \mathbb{R}^p$ とする、

とし、$p = d^2$ とする。 無制約部分問題(10)は、実ベクトルに対する典型的な最小化問題として考えることができる：

$$\min_{w2Rp} f(\boldsymbol{w}) + \lambda \| \boldsymbol{w} \|_1 、 \tag{14}$$

ここで、f（$\boldsymbol{w}$）＝l（W;X）＋ρ ／h（W）／^{2} ＋αh（W）である。 (15)

2

は目的の滑らかな部分である。我々の目標は、上記の問題を高精度で解き、以下のようにすることである。
$h(W)$は十分に抑えることができる。

$\lambda = 0$の特別な場合、非平滑項は消滅し、問題は単に制約のない平滑最小化となり、例えばL-BFGS [7]のような多くの効率的な数値アルゴリズムが利用可能である。非凸を扱うには、若干の修正 [28, Procedure 18.2] が必要である。

$\lambda > 0$のとき、問題は複合最小化となり、これは近接準ニュートン（PQN）法 [50]によっても効率的に解くことができる。各ステップkにおいて、重要な考え方は、平滑項の二次近似を通して降下方向を見つけることである：

$$\boldsymbol{d}_k = \arg\min_{d2Rp} \boldsymbol{g}^{\mathsf{T}}{}_k \boldsymbol{d} + \tfrac{1}{2}{}^{dT} B_k \, d + \lambda \| \boldsymbol{w}_k + \boldsymbol{d} \|_1 、 \tag{16}$$

ここで、$_{gk}$ は $f(\boldsymbol{w})$ の勾配、B_k はヘシアンの L-BFGS 近似である。なお、各座標jについて、問題（16）は以下のように与えられる閉形式の更新$\boldsymbol{d} \leftarrow \boldsymbol{d} + z\, e^?{}_j$ を持つ。

$$z^? = \arg\min_z \frac{1}{2} B \quad z^2 + (\boldsymbol{g} + (\boldsymbol{Bd}))z + \lambda | \boldsymbol{w} \quad + \boldsymbol{d} + z| = -c + S \checkmark c - \frac{b}{a}, \frac{\lambda}{a} . \tag{17}$$

さらに、B_k の低ランク構造により、座標更新のための高速計算が可能である。付録Bで説明するように、事前計算時間は$O(m^2 p + m^3)$のみで、m区pはL-BFGSのメモリサイズ、各座標更新はO（m）です。さらに、スパース性を利用しているため
正則化により、座標のサブグラディエント[50]に基づいてアクティブセットを積極的に縮小し、残りの次元を更新から除外することで、アルゴリズムをさらに高速化することができる。更新がアクティブセットSに制限されることで、$O(p)$の複雑さを持つ全ての依存関係
lは$O(|S|)$となり、大幅に小さくなる。したがって、L-BFGSの更新の全体的な複雑さは
$O(m^2 |S| + m^3 + m|S|T)$、ここでTは内部の反復回数で、通常$T = 10$です。

4.3 閾値処理

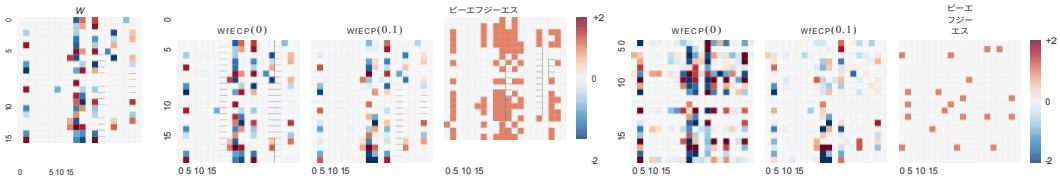
回帰問題では、ハードな閾値処理によって係数の推定値を後処理することで、偽の発見数を証明的に減らすことが知られている [46, 52]。これらの有望な結果に動機づけられ、我々は以下のようにエッジの重みを閾値□□□□る：(9)の定常点W を得た後、(9)の定常点W の上に$_{ECP}$
を固定閾値$\omega > 0$とし、絶対値でωより小さい重みを0とする。また、この戦略には
というのも，数値的な精度により，拡張ラグランジアン (9) の数値解は， $h(W_{ECP}) \leq$ □LI_3F5 を満たすが，機械精度に近い小さな公差ϵ（例：$\epsilon = 10^{-8}$）では，厳密に $h(W_{ECP}) = 0$ とならないからだ．しかし、$h(W_{ECP})$が明示的に
は、W_{ECP} の「DAG-ness」を定量化し（セクション3のデサイダータ（b）参照）、小さ

な閾値ωは、サイクルを誘発するエッジを除外するのに十分な値であると言える。

5 実験風景

我々は、本手法を greedy equivalent search (GES) [9, 31], PC algorithm [42], LiNGAM [38]と比較した。GESについては、Ramseyら[31]の高速貪欲探索（FGS）実装を使用した。PCとLiNGAMの精度はFGSとNOTEARSのいずれよりも著しく低かったため、ここではFGSに対する結果のみを報告する。これはスコアベース学習に関する過去の研究 [2] と一致しており、FGSがヒルクライミングやMMHC [44] などの他の技術を凌駕していることも示している。FGSが選ばれたのは、大規模な問題にも対応できる最先端のアルゴリズムであるためである。

簡潔にするため、ここでは実験の基本的なセットアップを概説する。すべてのパラメータの選択とより詳細な評価を含む実験セットアップの正確な詳細は、付録Eに記載されています。



(a) 真のグラフ　(b) エヌエックス　(c) $n = 20$で推定

図2：スケールフリーグラフ上のW_{ECP} のパラメータ推定値。アルゴリズム1の閾値設定ステップを追加しなくても、NOTEARSは真のグラフの一貫した推定値を生成しています。提案手法は、正則化なしでも大きなサンプルで重みを非常によく推定し、l_1 -正則化を導入しても不十分なサンプルで正確さを保つ。図1も参照してください。

各実験では、Erdös- Rényi (ER) と scale-free (SF) の2つのランダムグラフモデルからランダムグラフGを生成した。Gが与えられたとき、一様にランダムなエッジの重みを割り当てて、重み行列Wを得た。W が与えられたとき，$X = W^{\mathsf{T}} X + z \in \mathrm{R}^{\mathrm{d}}$ を3つの異なるノイズモデルからサンプリングした．ガウス（Gauss）、エクスポネンシャル（Exp）、ガンベル（Gumbel）。これらのモデルに基づいて、生成されたランダムデータセット $\in \mathrm{R}^{\mathrm{n}\times\mathrm{d}}$ は、$d \in \{10, 20, 50, 100\}$と$n \in \{20, 1000\}$の3つのモデルのいずれかに従って、行を i.i.d. 生成することによって生成される。FGSはDAGや重み行列の代わりにCPDAGを出力するため、比較を行う際には若干の注意が必要である。詳細は付録E.1を参照。

5.1　パラメータ推定

まず、*閾値処理を行わない*NOTEARSで得られた解の定性的な検討を行いました
を解くことで得られる重み行列W_{ECP} を可視化することで、（ECP）（$=\omega=0$）を実現する。これは図1（ER-2）と図2（SF-4）に示されています。重要なことは、本手法が真の重み行列Wの（経験的に）一貫したパラメータ推定値を提供することです。アルゴリズム1の最後の閾値のステップは、構造学習の精度を確保するためにのみ必要である。また、l_1 -regularizationが小さなn領域でいかに効果的であるかを示しています。

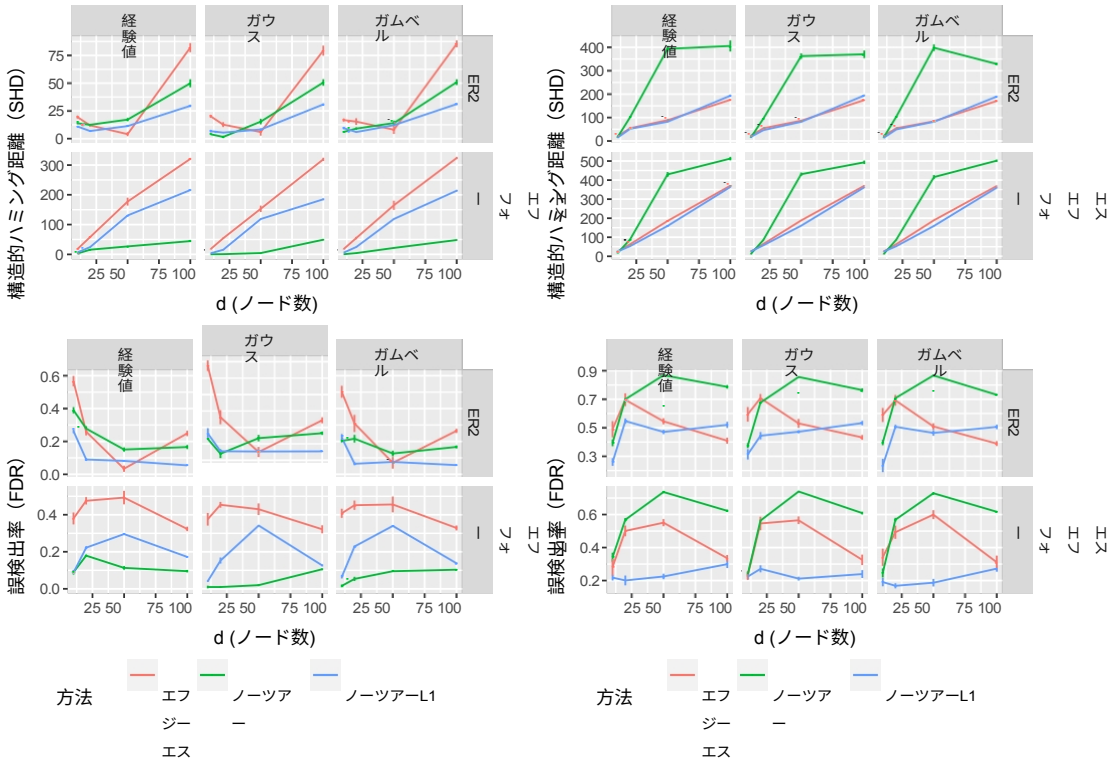
5.2　構造学習

次に、図3に示すような構造回復の方法を検討します。ここでは簡潔にするため、構造ハミング距離（SHD）の数値のみを報告するが、追加のメトリクスに関する完全な図と表は、付録にある。貪欲な手法に関するこれまでの研究と同様に、FGSはエッジの数が少ないときは非常に競争力があるが（ER-2）、エッジの数が控えめでも急速に悪化する（SF-4）。後者の領域では、NOTEARSは大きな改善を示した。これは評価した各メトリクスで一貫しており、ノード数dが大きくなるにつれてその差は大きくなる。また、我々のアルゴリズムは、ノイズの種類に関する特定の知識を活用することなく、各ノイズモデ

ル（Exp、Gauss、Gumbel）に対して一様に優れた性能を発揮することに注目。ここでも、l_1 -regularizerは、小さなnの設定に大きく役立っています。

5.3 厳密なグローバルミニマイザーとの比較

(3)で与えられる元のプログラムを解く本手法の能力を評価するために、GOBNILPプログラム[12, 13]を用いて(3)の正確な最小化子を求めた。これは、各ノードに対して可能なすべての親集合を列挙することを含むので、これらの実験は小さなDAGに限定される。しかし、これらの小規模な実験から、NOTEARSが実際に元の問題を解く際にどの程度の性能を発揮するのかについて、貴重な知見を得ることができた。実験では、$d = 10$のランダムグラフを生成し、$n = 20$のサンプル（高次元用）と$n = 1000$のサンプル（低次元用）を含む10のシミュレーションデータセットを生成しました。そして、本手法が返すスコアを、推定されたパラメータとともにGOBNILPが計算した厳密な大域最小化器と比較した。その結果を表1に示す。驚くべきことに、NOTEARSは局所最小化器を返すことだけを保証しているが、多くの場合、GOBNILPは局所最小化器を返さない。
は，偏差 $\mathbf{W_C} - W_G$ から明らかなように，大域的最小化器に非常に近い解を得ることができた．ということは
一般的な構造学習問題はNP困難であるが、我々がテストしたモデル（ERとSF）は高速に解けるように見えるが、最悪の場合、実行に指数関数的な時間がかかったり、局所最小値にはまるグラフが存在するのではないかと考えている。さらに、この問題は



(a) n = 1000の場合のSHD　　(b) n = 20のSHD

図3：真のグラフに対するSHDとFDRの観点から見た構造回復（低い方が良い）。行：ran-domグラフの種類、{ER,*SF*}-k = {Erdös-Rényi, scale-free} graphs with *kd* expected edges. 列：SEMのノイズタイプ．エラーバーは10回のシミュレーションによる標準誤差を表す。

表1：NOTEARSとグローバル最適解との比較。$\Delta(W_G, W^C) = F(W_G) - F(W^C)$.

n	λ	グラフ	$F(W)$	$F(W_G)$	$F(W^C)$	$F(W_{\mathrm{fECP}})$	$\Delta(W_G, W^C)$	$W^C - W_G$	$W - W_G$
20	0	ER2	5.11	3.85	5.36	3.88	-1.52	0.07	3.38
20	0.5	ER2	16.04	12.81	13.49	12.90	-0.68	0.12	3.15
1000	0	ER2	4.99	4.97	5.02	4.95	-0.05	0.02	0.40
1000	0.5	ER2	15.93	13.32	14.03	13.46	-0.71	0.12	2.95
20	0	エスエフフォー	4.99	3.77	4.70	3.85	-0.93	0.08	3.31
20	0.5	エスエフフォー	23.33	16.19	17.31	16.69	-1.12	0.15	5.08
1000	0	エスエフフォー	4.96	4.94	5.05	4.99	-0.11	0.04	0.29
1000	0.5	エスエフフォー	25.29	17.56	19.70	18.43	-2.13	0.13	4.34

*kd*が大きくなるにつれて難しくなる。それにもかかわらず、これは(8)の非凸性が実際には小さな問題であることを示す心強い証拠である。これらの問題をさらに調査することは、今後の研究に委ねられる。

5.4 リアルデータ

また、Sachs ら[33]が提供した実データセットでFGSとNOTEARSを比較した。このデータセットは、ヒト免疫系細胞におけるタンパク質とリン脂質の発現レベルの連続測定からなる（$n = 7466$ $d = 11, 20$ edges）。このデータセットは、既知の*コンセンサスネットワーク*、*つまり*生物学的コミュニティによって広く受け入れられている実験的注釈に基づくゴールドスタンダードネットワークが付属しているので、グラフィカルモデルにおける一般的なベンチマークとなるものである。我々の実験では、FGSはSHDが22の17個のエッジを推定したのに対し、NOTEARSはSHDが22の16個であった。

6 ディスカッション

我々は、連続最適化プログラムに基づき、データからDAGを学習する新しい手法を提案した。これは、DAGの離散空間上を探索し、結果として困難な最適化プログラムをもたらす既存のアプローチとは大きく異なる。また、得られたプログラムを定常的に解くための2つの最適化スキームを提案し、貪欲な等価探索などの既存の方法に対する優位性を説明した。重要なのは、各反復において局所的な更新（例えば一度に一つの辺）ではなく、大域的な更新（例えば一度に全てのパラメータ）を行うことにより、本手法はグラフの局所構造に関する仮定に依存することを避けることができることである。最後に、我々の手法の限界と、今後の研究の方向性について述べる。

まず、等式制約プログラム(8)が非凸プログラムであることをもう一度強調しておく必要がある。したがって、我々は*組み合わせ*最適化の難しさを克服したが、我々の定式化は依然として*非凸*最適化に関連する難しさを引き継いでいる。特に、ブラックボックスソルバーは、せいぜい(8)の定常点を見つけるのが精一杯である。しかし、厳密な手法を除いて、既存の手法もこの欠点に悩まされています。[2]NOTEARSの主な利点は、組合せ的な局所探索とは対照的に、*滑らかで グローバルな探索であり*、さらに探索は標準的な数値ソルバーに委ねられていることである。

第二に、現在の研究は、グラフ探索を導くために勾配ベースの数値ソルバーを利用するために、スコア関数の滑らかさに依存している。しかし、BDe [20]のような非平滑なスコアや離散的なスコアも考慮することは興味深い。Nesterovの平滑化[26]のような既成の技法は有用であるが、より詳細な調査は今後の課題である。

第三に、行列の指数の評価は$O(d^3)$であるため、本手法の計算量はノード数の3乗となりますが、疎な行列ではこの定数は小さくなっています。実際、これは我々が（1次ではなく）2次法を用いる重要な動機の1つです、
つまり、行列の指数計算の回数を減らすことができる。2次法を用いることで、各反復は1次法よりも大幅に進歩する。さらに、実際にはそれほど多くの反復（t □）は必要ないが、我々は最悪の場合の反復を確立していない。
の複雑さの結果である。5.3節の結果に照らし合わせると、以下のような例外的なケースがあると予想される。
収束が遅い。しかし、NOTEARSは、既存手法の難所として知られるインディグリーが大きい場合に、すでに既存手法を凌駕しています。これらのケースをより深く研究することは、今後の研究に委ねます。

最後に、我々の実験では、閾値設定に最適とは言えない固定値$\omega > 0$を選択した（セクション4.3）。明らかに、異なるノイズ対信号比やグラフタイプに適応する、データ駆動型のωの選択を見つけることが望ましいだろう。このような選択肢を研究することは、今後の興味深い方向性である。

コードは、`https://github.com/xunzheng/notears` で公開されています。

謝辞

貴重なご意見をいただいた匿名査読者の方々に感謝します。P.R.は、IIS-1149803、IIS-1664720によるNSFの支援に感謝する。E.X.とB.A.は、NIH R01GM114311, P30DA035778のサポートに感謝します。X.Z.は、Department of Health BD4BH4100070287、NSF IIS1563887、AFRL/DARPA FA87501720152の支援を受けたことを認める。

参考文献

[1]Al-Mohy, Awad H., & Higham, Nicholas J. 2009.A New Scaling and Squaring Algorithm for the Matrix Exponential.*SIAM Journal on Matrix Analysis and Applications*.

[2]アラガム，ブライオン，アンド ジュー，チング．2015.スパースガウスベイズネットワークの凹型罰則付き推定．*機械学習研究*, **16**, 2273-2328.

[3]アラガム，ブライオン，アミニ，アラシュA.，＆チュウ，チング．2016.罰則付き近傍回帰による有向無尽グラフの学習。*Submitted*, **arXiv:1511.08963**.

---

[2]GES [9]は、ある仮定のもとで極限*n ...1*において大域最小化器を求めることが知られている！*1 を見つけることが知られ*ているが、これは有限サンプルでは保証されない。

[4]Banerjee, Onureena, El Ghaoui, Laurent, & d'Aspremont, Alexandre.2008.多変量ガウスデータまたはバイナリデータに対するスパース最尤推定によるモデル選択.*機械学習研究*, **9**, 485-516.

[5] バラバシ，アルベルト＝ラースロー，＆アルベルト，レカ．1999.ランダムネットワークにおけるスケーリングの創発.
Science, **286**(5439), 509-512.

[6]Bouckaert, Remco R. 1993.最小記述長原理を用いた確率的ネットワーク構築.*推論と不確実性への記号的・定量的アプローチに関するヨーロッパ会議*.Springer, pp.41-48.

[7]Byrd, Richard H., Lu, Peihuang, Nocedal, Jorge, & Zhu, Ciyou.1995.境界制約付き最適化のための限定メモリ・アルゴリズム.*SIAM Journal on Scientific Computing*.

[8]チッカリング，デイビッド・マックスウェル．1996.ベイジアンネットワークの学習はNP-completeである。In *Learning from data*.Springer.

[9]チッカリング，デイビッド・マックスウェル．2003.貪欲な探索による最適な構造同定.*機械学習研究*, **3**, 507-554.

[10] チッカリング，デイビッド・マックスウェル，＆ヘッカーマン，デイビッド．1997.隠れ変数を持つベイジアンネットワークの周辺尤度の効率的な近似値。*機械学習*, **29**(2-3), 181-212.

[11] Chickering, David Maxwell, Heckerman, David, & Meek, Christopher.2004.ベイジアンネットワークの大標本学習はNP-hardである。*機械学習研究*, **5**, 1287-1330.

[12] Cussens, James.2012.*カッティングプレーンによる*ベイジアンネットワークの学習。

[13] Cussens, James, Haws, David, & Studeny`, Milan.2017.ベイズネットワーク構造学習におけるスコア等価性の多面体的側面。*Mathematical Programming*, **164**(1-2), 285-324.

[14] エリス、バイロン、ウォン、ウイング・フン。2008.実験データから因果関係ベイジアンネットワーク構造を学習する。*アメリカ統計学会誌*, **103**(482).

[15] Friedman, Jerome, Hastie, Trevor, & Tibshirani, Robert.2008.Graphical Lassoによるスパース逆共分散推定。*Biostatistics*, **9**(3), 432-441.

[16] Fu, Fei, & Zhou, Qing.2013.実験介入を伴うスパース因果ガウスネットワークの学習：正則化と座標降下.*アメリカ統計学会誌*, **108**(501), 288-300.

[17] Gámez, José A, Mateo, Juan L, & Puerta, José M. 2011.ヒルクライミングによるベイジアンネットワークの学習：近傍の漸進的な制限に基づく効率的な方法。*データマイニングと知識発見*, **22**(1-2), 106-148.

[18] Gu, Jiayang, Fu, Fei, & Zhou, Qing.2018.離散データからの有向非周期グラフのペナルティ付き推定(Penalized Estimation of Directed Acyclic Graphs From Discrete Data).*Statistics*

and Computing, **DOI: 10.1007/s11222-018-9801-y**.

[19] ハラリー，フランク＆マンベル，ベネット。1971.グラフのサイクルの数について。*Matematicky` c˘asopis*.

[20] Heckerman, David, Geiger, Dan, & Chickering, David M. 1995.ベイジアンネットワークの学習：知識と統計データの組み合わせ。*機械学習*, **20**(3), 197-243.

[21] Hsieh, Cho-Jui, Sustik, Mátyás A, Dhillon, Inderjit S, & Ravikumar, Pradeep.2014.QUIC: quadratic approximation for sparse inverse covariance estimation.*機械学習研究*, **15**(1), 2911-2947.

[22] コラー、ダフネ、＆フリードマン、ニル。2009.*確率的グラフィカルモデル：原理と技法*.MIT press.

[23] Kuipers, Jack, Moffa, Giusi, & Heckerman, David.2014.ガウス型有向非周期グラフモデルのスコアリングに関する補遺。*統計学年鑑*』1689-1691頁。

[24] Loh, Po-Ling, & Bühlmann, Peter.2014.Inverse Covariance Estimation による Linear Causal Net- works の高次元学習。*機械学習研究誌*、**15**、3065-3105.

[25] ネミロフスキー，アルカディ．1999.最適化II：非線形連続最適化のための標準的な数値計算法.

[26] ネステロフ、ユリイ。2005.非平滑関数の滑らかな最小化.*Mathematical Program- ming*.

[27] Niinimäki, Teppo, Parviainen, Pekka, & Koivisto, Mikko.2016.部分順序のサンプリングによるベイジアンネットワークの構造発見。*機械学習研究*, **17**(1), 2002-2048.

[28] Nocedal, Jorge, & Wright, Stephen J. 2006.*数値的最適化*

[29] Ott, Sascha, & Miyano, Satoru.2003.生物学的コンストレイントを用いた最適な遺伝子ネットワークの発見。*ゲノムインフォマティクス*, **14**, 124-133.

[30] パク・ヨンウン、＆クラビャン、ディエゴ。2017.トポロジカルオーダーを介したベイジアンネットワーク学習.*ジャーナル・オブ・マシーンラーニング・リサーチ*。

[31] ラムゼイ、ジョセフ、グリモア、マデリン、サンチェス-ロメロ、ルーベン、＆グリモア、クラーク。2016.A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high dimensional graphical causal models, with an application to functional magnetic resonance images.*International Journal of Data Science and Analytics*, pp.1-9.

[32] Robinson, Robert W. 1977.ラベルのない非周期的なディグラフをカウントする。*組合せ論的数学において*
V.スプリンガー

[33] Sachs, Karen, Perez, Omar, Pe'er, Dana, Lauffenburger, Douglas A, & Nolan, Garry P. 2005. マルチパラメータ単一細胞データから導き出された原因タンパク質シグナル伝達ネットワーク。*Sci- ence*, **308**(5721), 523-529.

[34] Scanagatta, Mauro, de Campos, Cassio P, Corani, Giorgio, & Zaffalon, Marco.2015.数千の変数を持つベイジアンネットワークを学習する。In *Advances in Neural Information Processing* Systems. pp.1864-1872.

[35] Scanagatta, Mauro, Corani, Giorgio, de Campos, Cassio P, & Zaffalon, Marco.2016.数千の変数を持つ木幅境界ベイジアンネットワークの学習.In *Advances in Neural Information Processing* Systems. pp.1462-1470.

[36] Schmidt, Mark, Niculescu-Mizil, Alexandru, & Murphy, Kevin.2007.L1-正則化パスを用いたグラフィカルモデル構造の学習。*AAAI*, vol.7. pp.1278-1283.

[37] Schmidt, Mark, Berg, Ewout, Friedlander, Michael, & Murphy, Kevin.2009.単純な制約を持つ高価な関数を最適化する：限定メモリ投影型準ニュートン・アルゴリズム.*人工知能と統計学*』pp.456-463.

[38] 清水翔平、Hoyer, Patrik O、Hyvärinen, Aapo, & Kerminen, Antti.2006.因果関係発見のための線形非ガウス型非周期モデル.*機械学習研究*, **7**, 2003-2030.

[39] Silander, Tomi, & Myllymaki, Petri.2006.大域的に最適なベイジアンネットワーク構造を見つけるための簡単なアプローチ。*第22回人工知能における不確実性会議予稿集*.

[40] Singh, Ajit P, & Moore, Andrew W. 2005.動的計画法による最適なベイジアンネットワークの発見。

[41] スピルテス、ピーター、グリモア、クラーク。1991.疎な因果グラフを高速に復元するためのアルゴリズム。
社会科学コンピューターレビュー, **9**(1), 62-72.

[42] Spirtes, Peter, Glymour, Clark, & Scheines, Richard.2000.*因果関係、予測、探索*。第81巻。The MIT Press.

[43] Teyssier, Marc, & Koller, Daphne.2005.順序に基づく探索：ベイジアンネットワークを学習するためのシンプルで効果的なアルゴリズム。In *Uncertainty in Artifical Intelligence (UAI)*.

[44] Tsamardinos, Ioannis, Brown, Laura E, & Aliferis, Constantin F. 2006.ベイジアンネットワーク構造学習アルゴリズム（Max-min Hill- climbing Bayesian Network Structure Learning algorithm）。*機械学習*, **65**(1), 31-78.

[45] van de Geer, Sara, & Bühlmann, Peter.2013. l_0 -penalized maximum likelihood for sparse directed acyclic graphs.*統計学年報*, **41**(2), 536-567.

[46] Wang, Xiangyu, Dunson, David, & Leng, Chenlei.2016.No penalty no tears：高次元線形モデルにおける最小二乗法。In *International Conference on Machine Learning*. pp.1814- 1822.

[47] Xiang, Jing, & Kim, Seyoung.2013.連続変数の疎なベイジアンネットワーク構造を学習するためのA* Lasso.In *Advances in Neural Information Processing Systems*. pp.2418-2426.

[48] Yuan, Ming, & Lin, Yi.2007.ガウスグラフィカルモデルにおけるモデル選択と推定. *Biometrika*, **94**(1), 19-35.

[49] Zhang, Bin, Gaiteri, Chris, Bodea, Liviu-Gabriel, Wang, Zhi, McElwee, Joshua, Podtelezhnikov, Alexei A, Zhang, Chunsheng, Xie, Tao, Tran, Linh, Dobrin, Radu, *et al.*2013.Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.*Cell*, **153**(3), 707-720.

[50] Zhong, Kai, Yen, Ian En-Hsu, Dhillon, Inderjit S, & Ravikumar, Pradeep K. 2014 年 .Proximal quasi-Newton for computationally intensive l1-regularized m-estimators.In *Advances in Neural Information Processing Systems*. pp.2375-2383.

[51] Zhou, Qing.2011.マルチドメインサンプリングとベイジアンネットワークの構造推論への応用(Multi-Domain Sampling With Applications to Structural Inference of Bayesian Networks). *アメリカ統計協会誌*, **106**(496), 1317-1330.

[52] Zhou, Shuheng.2009.高次元の変数選択と統計的推定のための閾値処理手続き.In *Advances in Neural Information Processing Systems*. pp.2304-2312.