電気通信大学情報理工学域
I類 (情報系) 情報数理工学プログラム卒業論文

統計的因果推論のための
ベイジアンネットワーク分類器による傾向スコア推定

2025年1月30日

コンピュータサイエンスプログラム

学籍番号 2110293

佐久間 理奈

指導教員 植野 真臣

令和 6 年度 情報数理工学プログラム卒業論文概要

令和 3 年度　入学	学籍番号　2110293
指導教員　植野 真臣	氏名　佐久間 理奈
題目　統計的因果推論のためのベイジアンネットワーク分類器による傾向スコア推定	

概要

因果推論は, 特定の介入や行動が結果に与える影響を推定するための手法であり, 医療, 経済学, 社会科学などの多くの分野で広く利用されている. 無作為化比較試験が実施できる場合は, 結果と介入の有無が独立となるため, 介入を受けた群 (介入群) と介入を受けなかった群 (対照群) を比較することで介入効果を推定できる. 一方, 無作為化比較試験が難しい場合は, 介入群と対照群で共変量の分布に違いがあるため, 傾向スコアにより共変量の分布を調整する必要がある. しかし, 傾向スコアの一般的な推定手法であるロジスティック回帰は, 傾向スコアの logit が共変量の線形関数で表現できない場合, 傾向スコアの推定に一致性が無くなる. さらに, 機械学習手法を用いた傾向スコア推定法は, 漸近的に真の傾向スコアを推定できず, 結果の解釈が難しくなるという課題がある.

近年, 真のモデルに依存せず, 真の分類確率に漸近収束するベイジアンネットワーク分類器学習法が提案されている. そこで本研究では, この手法を用いた新しい傾向スコア推定法を提案する. 提案手法は, 変数間の因果関係を視覚的に表現でき, その推定値は真の傾向スコアに漸近的に一致する.

提案手法の有効性を示すため, マルコフネットワークを用いたシミュレーションと, 実データセット jobs と twins を用いた実験を行い, 従来手法と比較する. これにより, 提案手法は傾向スコアの推定精度を向上させることで, 高精度に介入効果を推定できることを示す.

1　まえがき

反事実の仮定をおいた Rubin 因果モデル [1] は, 介入を受けた場合の結果と介入を受けなかった場合の結果を比較することで因果推論を行う統計的手法であり, 医療, 経済学, 機械学習などの分野で高い注目を集めている [2, 3, 4]. 例えば, 冠動脈疾患が疑われる, または既知の評価を受けている患者におけるアスピリン使用と全死亡率の関係を調査し, アスピリン使用が全死亡率に与える影響を評価したもの [2] や, 労働トレーニングプログラムの介入効果の推定 [3] など, 多くの応用が存在する. 因果推論の中心にある課題は, 対象が介入を受けた場合と受けなかった場合の両方の結果を同時に観測することができないという反実仮想の推論が困難なことである.

無作為化比較試験が実施できる場合は, 結果と介入の有無が独立となるため, 介入を受けた群 (介入群) と介入を受けなかった群 (対照群) を比較することで因果効果を推定できる. しかし, 無作為化比較試験が実施できない観察研究の場合は, 介入群と対照群で共変量の分布に違いがあるため, 因果効果の推定にバイアスが生じる.

この問題を解決するために Rosenbaum and Rubin[5] は, 対象が介入を受ける確率である傾向スコアを用いた介入群と対照群の共変量調整法を提案した. 具体的には, 傾向スコアを用いたマッチング [5] や傾向スコアの逆数による重み付け法 [6] によって, 観察研究においても, 無作為化比較試験と同様に介入効果の推定が可能になる手法を提案した.

しかし, 傾向スコアの正確な推定は容易ではない. 特に, 一般的な推定手法であるロジスティック回帰モデルを用いた傾向スコアの推定 [5] は, 傾向スコアの logit が共変量の線形関数で表現できない場合, 傾向スコアの推定に一致性が無くなる [7]. 結果的に介入効果の推定にバイアスが生じることがある. また, 実データでは, 共変量の次元が非常に高い場合や, 複雑な依存構造を持つ場合が多く, そのような場合, ロジスティック回帰では傾向スコアの推定が困難である. これにより, 介入効果の正確な推定が困難になり, 誤った結論に至るリスクが生じる.

近年, 傾向スコア推定のためのより高度なアプローチとして, 勾配ブースティング [8] やニューラルネットワーク [9], 決定木 [9, 10] といった機械学習手法の利用が注目されている. これらの手法は非線形性や高次元の変数を扱う能力に優れており, 従来のロジスティック回帰モデルに比べより正確な傾向スコア推定が期待される. しかし, 漸近的に真の確率を推定することができないという課題がある. また, モデルの解釈性の低さから結果の解釈が難しくなるという問題もある [10]. 因果推論の文脈においては, 説明可能性や結果の透明性が求められるケースが多く, これが機械学習手法の応用を難しくしている要

因の一つとなっている.

近年, これらの機械学習手法の問題を解決できる新しいモデルが提案されている. Sugahara ら [11] は, 真の同時確率を表現可能な構造のうち, 目的変数パラメータ数を最小化する新しいベイジアンネットワーク分類器 (Bayesian Network Classifier: BNC) 構造学習法を提案した. この手法は, 真のモデルがベイジアンネットワークに従うか否かに関わらず, 真の分類確率に漸近的に一致することが特徴である. さらに, この手法は自動的に変数選択をすることが可能であり, 変数間の条件付き独立性をグラフィカルモデルとして可視化することができる.

そこで, 本研究は, Sugahara ら [11] のベイジアンネットワーク分類器を用いた新しい傾向スコア推定法を提案する. 提案手法では, 介入の有無を目的変数とし, 共変量と結果変数を説明変数として設定し, 分類に影響を与える目的変数のパラメータ数を最小化することで, 真の傾向スコアに漸近的に一致する BNC を学習する. 提案手法は真のモデルがベイジアンネットワークに従わない場合でも真の傾向スコアを漸近的に推定できる. さらに因果関係を視覚的に表現することで結果を理解しやすくなり, データから構造を学習し, 状況に応じた最適なモデルを構築可能であるという利点がある.

本研究の目的は, このベイジアンネットワーク分類器を用いて, 従来の手法と比較して因果推論における傾向スコア推定の精度を向上させ, より正確に因果効果を推定することである. 特に, ロジスティック回帰 [5], 勾配ブースティング [7] やニューラルネットワーク [9] といった既存手法との比較を通じて, 提案手法の優位性を検証する. 本研究では, まずマルコフネットワークを用いたシミュレーションデータを用いて提案手法の性能を評価し, 次に実データセット jobs と twins を使用した実験を行う. これらの実験を通じて, 提案手法が他の手法と比較して高精度に因果効果を推定できることを示す.

2 因果推論

2.1 因果推論の基礎概念

因果推論は, 特定の介入や行動が結果に与える影響を推定するための手法であり, 医療, 経済学, 社会科学などの多くの分野で広く利用されている [2, 3]. 例えば医療分野では, アスピリンを処置した際の冠動脈疾患に対する有効性 [2] や心筋梗塞患者に対する救急治療の方法による予後の影響 [12] に対して因果推論が使われ,　経済学においては, 特定の社会政策が経済成長や失業率に与える影響を推定するために利用される [13]. 因果推論の基本的な考え方は, ある対象に対して介入が行われた場合と行われなかった場合の反実仮想

的結果を比較することにある.

対象が介入を受けた場合の結果変数を y_1, 介入を受けなかった場合の結果変数を y_0 とする. また, 変数 z は対象が介入を受けたとき $z = 1$, 介入を受けなかったとき $z = 0$ を取るとする. $z = 1$ となる対象の集合を介入群, $z = 0$ となる対象の集合を対照群とすると, 介入群, 対照群の平均介入効果（Average Treatment Effect : ATE）は以下のように定義される [1].

$$ATE = E(y_1 - y_0) = E(y_1) - E(y_0) \tag{1}$$

ここで, $E(y_1)$ はすべての対象が介入を受けたときの結果の期待値, $E(y_0)$ はすべての対象が介入を受けなかったときの結果の期待値を表す.

しかし, 同一の対象では y_1 と y_0 を同時に観測することはできない. 例えば, ある患者が薬を服用した場合の健康状態と, 同じ患者が薬を服用しなかった場合の健康状態を同時に観察することはできない.

2.2 無作為化比較試験

無作為化比較試験（Randomized Controlled Trial: RCT）は, 因果推論において最も確実な方法とされていて, 被験者を無作為に介入群と対照群に割り振り, 介入の効果を評価する. 無作為化により, 介入群と対照群の間で共変量の分布が均一化されるため, 介入が結果に与える純粋な影響を推定することが可能である. 例えば, 新しい薬剤の効果を評価するために患者を無作為に 2 つのグループに分け, 一方に薬を投与し, もう一方にプラセボを投与する実験デザインが典型的な RCT である. このように無作為化することで, 患者の年齢, 性別, 健康状態などの共変量が結果に影響を与えないように調整され, 薬剤の純粋な効果を正確に推定することができる. また, 被験者がどちらの群に割り振られたか分からないようにすることでバイアスを防ぐ. 無作為化比較試験では, $(y_1, y_0) \perp z$ が成り立つため,

$$E(y_j) = E(y_j \mid z = j), \quad (j = 1, 0) \tag{2}$$

となる. したがって,

$$ATE = E(y_1 - y_0) = E(y_1 \mid z = 1) - E(y_0 \mid z = 0) \tag{3}$$

となり, 介入群での結果の期待値と対照群での結果の期待値の差で因果効果を求めることができる. RCT は理論上理想的な手法だが, 現実の場面では様々な理由で実施が困難な場

合が多くある．例えば, 倫理的な理由で特定の介入を一部の人々にのみ提供することができない場合や, コストがかかるために無作為化試験が不可能な場合である．その場合, 介入群と対照群で共変量の分布が一致していないため,

$$ATE \neq E(y_1 \mid z = 1) - E(y_0 \mid z = 0) \tag{4}$$

であり, ATE を正確に計算することができない.

2.3 観察データにおけるバイアス

無作為化比較試験が実施できない状況では, 観察データに基づいて因果推論を行う必要がある．しかし, 観察データを用いる場合, 介入群と対照群の共変量の分布が一致していないことが一般的である．これは, 介入の効果が過大評価または過小評価されるバイアスが生じる原因となる．このバイアスを調整し, 因果効果を正確に推定するためには, 共変量の調整が必要である.

共変量の調整には主にマッチング [5] が用いられる．マッチングでは, 介入群と対照群で共変量の値が同じになるペアを作り, ペア間で比較を行う．例えば, 年齢, 性別, 健康状態が同じ投薬を行った群とプラセボを投与した群の患者をペアにして比較を行い, 薬の効果を推定する．しかし, マッチングにおいては共変量に連続量が存在する場合, 完全マッチングができない．また, 共変量の数が大きいと実行が難しく, 共変量の分布が介入群と対照群で重ならない部分がある時にはこの手法は用いることができない．これらの問題を解決する新しい調整法として近年, 傾向スコアによる手法が注目を集めている.

3 傾向スコア

3.1 傾向スコアの定義

傾向スコア [5] は, 無作為化試験ができない観察データにおいて, 介入群と対照群の共変量のバランスを調整するための手法で, Rosenbaum and Rubin（1983）によって提案された．傾向スコア $e(x)$ は, 特定の共変量 x を持つ対象が介入を受ける確率を示すスコアであり, 以下のように定義される.

$$e(x) = p(z = 1|y_1, y_0, x) = p(z = 1|x) \tag{5}$$

ここで, x は共変量のベクトルである. このとき,

$$(y_1, y_0) \perp z \mid e(x) \tag{6}$$

が成り立つ [5] ため,

$$E(y_j \mid x) = E(y_j \mid e(x)) \quad (j = 0, 1) \tag{7}$$

が成り立つ. したがって, 共変量 x を一次元の値として利用することが可能になる. 傾向スコアを用いることで, 介入群と対照群の共変量の分布を一致させ, 共変量によるバイアスを排除することができる. この手法を用いると, 無作為化比較試験と同様に, 共変量の影響を排除した形で介入効果を推定できるため, 観察データでも因果推論が適用可能となる.

3.2 傾向スコアの利用方法

傾向スコアは, 主に以下の方法で利用される.

1. **傾向スコアマッチング（Propensity Score Matching : PSM）[5]**
   介入群と対照群の個体を傾向スコアでマッチングさせ, 同じ傾向スコアを持つ対象同士を比較することで, 共変量によるバイアスを除去する.
2. **逆確率重み付け（Inverse Probability Weighting : IPW）[6]**
   各対象に対して傾向スコアに基づいた重みを付け, 介入群と対照群の共変量の分布を調整する.
3. **傾向スコアによる層別化 (Stratification or Subclassification)[14]**
   傾向スコアに基づいてサンプルをいくつかの層に分割し, 各層ごとに介入効果を推定する. 通常, 層の数は 5〜10 程度で各層内で共変量のバランスが取れていると仮定する.
4. **傾向スコア加重回帰 (Propensity Score Weighted Regression)[15]**
   傾向スコアを重みとして使用し, 重み付けされたデータを用いて回帰分析を行い, 介入効果を推定する. この方法は共変量調整によるバイアスの削減と回帰モデルの柔軟性を組み合わせている.
5. **ダブルロバスト推定 (Doubly Robust Estimation)[16]**
   傾向スコアと介入の有無に応じた結果を予測するモデル (アウトカムモデル) の両方を同時に用いることで, どちらか一方のモデルが正しく指定されていれば介入効果のバイアスが調整される. これにより, より信頼性の高い因果推定が可能となる.

本研究では, 最も一般的である IPW を使って ATE 推定を行う [17, 18]. IPW では傾向スコアに基づいて介入群と対照群の各対象に対して重みを付け, 全体のバランスを取ることで因果効果の推定が行われる. 以下の式で ATE の推定を行う.

$$ATE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_{1i}}{e(x_i)}z_i - \frac{y_{0i}}{1-e(x_i)}(1-z_i)\right) \tag{8}$$

ここで, N は対象全体の数, y_{1i} は対象 i が介入を受けた場合の結果, y_{0i} は対象 i が介入を受けなかった場合の結果, x_i は対象 i の共変量, z_i は対象 i が介入を受けたかどうかを示す指標を表す.

ここで, IPW により ATE を推定するためには, 傾向スコアをデータから推定する必要がある. Drake[19] で示されたように傾向スコアの推定が誤っていると, 推定される ATE に大きなバイアスが生じる可能性があるため, 高精度な傾向スコア推定法が必要になる.

3.3 傾向スコアの推定方法

傾向スコアの推定には通常, ロジスティック回帰（Logistic Regression : LR）[5] が使われる. 推定方法は以下のとおりである.

1. **目的変数と説明変数の設定**：
   - 目的変数は, 介入の有無（例：治療を受けたかどうか）
   - 説明変数は, 共変量（例：年齢、性別、既往歴など）
2. **ロジスティック回帰モデルの構築**：
   ロジスティック回帰分析を用いて説明変数から目的変数を予測するモデルを構築する. ロジスティック回帰モデルは以下のように表される.

$$\hat{e}(x) = \frac{1}{1+\exp^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{9}$$

   ここで $\hat{e}(x)$ は介入を受ける確率(傾向スコア)の推定値, β_0 は定数項, $\beta_1, \beta_2, \ldots, \beta_n$ は各説明変数の回帰係数, $x_1, x_2, \ldots, x_n$ は説明変数である.
3. **傾向スコアの算出**：
   ロジスティック回帰モデルを用いて、各対象の傾向スコア（介入を受ける確率）を算出する.

しかし, 傾向スコアの logit が共変量の線形関数で表現できない場合, 傾向スコアの推定

に一致性が無くなる [7]. logit は以下の式である.

$$\text{logit}(\hat{e}(x)) = \log \frac{\hat{e}(x)}{1-\hat{e}(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{10}$$

また, 高次元の共変量を含む場合, ロジスティック回帰による傾向スコアの推定精度が低下する問題がある.

近年, 傾向スコア推定のためのより高度なアプローチとして, 勾配ブースティング [8] やニューラルネットワーク [9], 決定木 [9, 10] といった機械学習手法の利用が注目されている. 以下に、代表的な機械学習手法をいくつか挙げる.

1. **勾配ブースティング（Gradient Boosting）[8]**
   勾配ブースティングは決定木をベースにしたアンサンブル学習手法で, 傾向スコア推定においてもその高いパフォーマンスが認められている. 勾配ブースティングは弱い予測器（浅い決定木）を複数組み合わせることで, 強力な予測器を作り上げる. この方法は非線形なデータや高次元データに対しても高い推定精度を発揮するが, 繰り返しのステップでモデルを改善するため計算量が多くなる問題がある.
2. **ニューラルネットワーク（Neural Network）[9]**
   ニューラルネットワークは非線形な関係をモデル化する能力に優れていて, 特に高次元データに対して有効である. 複数の層を持つニューラルネットワークはデータ内の複雑な依存関係を捉え, 傾向スコア推定にも応用されている. しかし, モデルの解釈が難しくなる点や過学習のリスクがある.
3. **決定木（Decision Tree）[9, 10]**
   決定木は, データを条件分岐に基づいて階層的に分類し, 介入を受ける確率を推定する手法である. 決定木は解釈が容易であり, モデルの可視化が可能であるため, 傾向スコア推定においても利用されている. 特に, 分類回帰木（CART: Classification and Regression Trees）は, 観察データの構造を反映する柔軟な手法として知られている.

これらの手法は非線形性や高次元の変数を扱う能力に優れており, 従来のロジスティック回帰モデルに比べより正確な傾向スコア推定が期待される. しかし, 漸近的に真の確率を推定することができないという課題がある. また, モデルの解釈性の低さから結果の解釈が難しくなるという問題もある [10]. そのため, 精度の高い推定が得られたとしても, 因果推論の文脈においては, 説明可能性や結果の透明性が求められるケースが多く, これが機械学習手法の応用を難しくしている要因の一つとなっている.

4 ベイジアンネットワーク分類器を用いた傾向スコア推定法の提案

これまでに真のモデルによらず推定値が漸近一致性を持ち, 自動的な変数選択が可能であるベイジアンネットワーク分類器 (Bayesian Network Classifier: BNC) が提案されている. そこで, 本論文ではこのベイジアンネットワーク分類器を用いた新しい傾向スコア推定法を提案する.

4.1 目的変数パラメータ数を最小にするベイジアンネットワーク分類器

ベイジアンネットワーク分類器は, ベイジアンネットワーク (Bayesian Network: BN) の一つであり, 確率変数をノードとし, ノード間の依存関係をエッジで表す非循環有向グラフ (Directed Acyclic Graph: DAG) G と, 各ノードの親ノード集合を所与とした条件付き確率パラメータ集合 Θ で表現される [20, 21, 22]. BNC では, G において一つのノードを目的変数, その他のノードを説明変数として扱い, 目的変数は親変数を持たないとする.

今, G は離散確率変数集合 $\mathbf{V} = \{X_0, X_1, \ldots, X_n\}$ の各変数をノードとして持つとする. また, 各変数 X_i は r_i 個の状態集合 $\{1, \ldots, r_i\}$ から一つの値を取るとし, X_i が状態 k を取るとき, $X_i = k$ と書く. さらに, X_0 を目的変数, $X_1, \ldots, X_n$ を説明変数とする. このとき, BNC では, 同時確率分布を条件付き確率の積に分解して以下のように表せる.

$$P(X_0, X_1, \ldots, X_n \mid G) = \prod_{i=0}^{n} P(X_i \mid \mathbf{Pa}(X_i, G), G). \tag{11}$$

ここで, $\mathbf{Pa}(X_i, G)$ は X_i の G における親変数集合を表す. また, 説明変数のデータ $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$ を得たとき, 目的変数の推定値 $\hat{c}$ は次式で表される.

$$\hat{c} = \underset{c \in \{1, \ldots, r_0\}}{\arg\max}\ P(c \mid \mathbf{x}, G, \Theta). \tag{12}$$

$\hat{c}$ を得るためには, 条件付き確率パラメータ集合 Θ をデータから推定する必要がある. 今, θ_{ijk} を $\mathbf{Pa}(X_i, G)$ が j 番目のパターンをとったとき ($\mathbf{Pa}(X_i, G) = j$ と書く) に, $X_i = k$ となる条件付き確率パラメータとする. ここで, 条件付き確率パラメータ集合 Θ は θ_{ijk} を用いて $\Theta = \bigcup_{i=0}^{n} \bigcup_{j=1}^{q_i} \bigcup_{k=1}^{r_i} \{\theta_{ijk}\}$ で定義される. ここで, q_i は $\mathbf{Pa}(X_i, G)$ の取りうるパターン数であり, $q_i = \prod_{l: X_l \in \mathbf{Pa}(X_i, G)} r_l$ で定義される. θ_{ijk} の推定法には, そ

の期待値である Expected A Posteriori (EAP) が最も良く用いられる. サンプルが N 個あるデータが得られた時の EAP は条件付き確率パラメータの事前分布にディリクレ分布を仮定すると以下で表される [23].

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}}. \tag{13}$$

ここで, N_{ijk} は $X_i = k$ かつ $\mathbf{Pa}(X_i, G) = j$ となる頻度を表す. また, α_{ijk} はディリクレ事前分布のハイパーパラメータを表し, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ である.

BNC の条件付き確率パラメータを推定するためには, 構造 G をデータから推定する必要があり, この問題を BNC の構造学習という. Sugahara ら [11] は, 分類確率のみを効率的に推定する BNC として, 以下で定義される目的変数パラメータ数 (Number of Class variable Parameters: NCP) を最小にして, 真の同時確率を表現可能な構造の学習法を提案した.

$$NCP(G) = \sum_{i=0}^{n} NCP_i(\mathbf{Pa}(X_i, G)). \tag{14}$$

ここで, $i = 0 \vee X_0 \in \mathbf{Pa}(X_i, G)$ のとき $NCP_i(\mathbf{Pa}(X_i, G)) = (r_i - 1)q_i$ であり, それ以外のとき $NCP_i(\mathbf{Pa}(X_i, G)) = 0$ である.

この手法は, 真のモデルがベイジアンネットワークに従うか否かに関わらず, 真の分類確率に漸近的に一致する構造を学習できる. また, 自動的に変数選択をすることが可能であり, 事前の変数選択が不要であるという利点がある. Sugahara ら [11] は比較実験を行い, この手法が BNC の中で最も高精度であることを示している.

4.2 目的変数パラメータ数最小 BNC を用いた傾向スコア推定

提案手法の傾向スコア推定方法は以下のとおりである.

1. **目的変数と説明変数の設定**：
   - 目的変数は, 介入の有無（例：治療を受けたかどうか）
   - 説明変数は, 共変量（例：年齢、性別、既往歴など）と結果変数
2. **モデルの構築**：
   - データから Sugahara ら [11] の手法で構造を学習する
   - 学習した構造から式 (13) を用いて条件付確率パラメータを推定する
3. **傾向スコアの推定**：
   以下の式で傾向スコアの推定を行う.

$$\begin{aligned}\hat{e}(x \mid G, \Theta) &= p(z = 1 \mid x, G, \Theta) \\ &= p(z = 1 \mid x_1, x_2, \ldots, x_n, G, \Theta) \\ &= \frac{p(z = 1, x_1, x_2, \ldots, x_n \mid G, \Theta)}{p(x_1, x_2, \ldots, x_n \mid G, \Theta)}\end{aligned} \tag{15}$$

ここで $\hat{e}(x \mid G, \Theta)$ は介入を受ける確率（傾向スコア）の推定値, $x_1, x_2, \ldots, x_n$ は説明変数, 学習したベイジアンネットワーク分類器の構造を G, 条件付き確率パラメータ集合を Θ とする.

提案手法の最大の特徴は, 真のモデルがベイジアンネットワークに従わない場合でも真の傾向スコアを漸近的に推定できる点にある.

さらに, 提案手法であるベイジアンネットワーク分類器は, ノード間の依存関係を視覚的に示すことができるため, モデルの解釈性が高い. また, 複数の共変量間の複雑な依存構造を直接的にモデル化することで, 共変量の次元性や非線形性を正確に捉えることができる. したがって, 傾向スコアの推定において, 従来の線形モデルに比べてより高精度な推定が期待される.

5 評価実験

この章では, 提案手法の有効性を検証するために, マルコフネットワーク (MN) のシミュレーションデータを用いた評価実験と実データセット jobs[3, 24, 25], twins[26, 27, 28] を用いた 3 つの評価実験を行う. シミュレーション実験と jobs データを用いた実験は以下の 5 つの手法で比較を行う.

- ロジスティック回帰 (Rosenbarum ら [5], 以降 LR と表記)
- Boosted CART (Lee ら [29], 以後 BOOST と表記)
- ニューラルネットワーク (Setoguchi ら [9], 以後 NN と表記)
- 制約のない一般的なベイジアンネットワーク (以後, GBN と表記)
- 提案手法

twins データは変数数が多く, GBN では 24 時間以上の計算時間が必要になる. したがって, twins データでは, GBN 以外の 4 つの手法で比較を行った. ロジスティック回帰は Python で実装し, GBN と提案手法については C++ で実装した. Boosted CART は [29] の論文を参考にした. R 言語で実装し, twang パッケージを使用した. McCaffrey ら [8] によって推奨されたパラメータを使用し、20,000 回の反復回数と 0.0005 の縮小パラメー

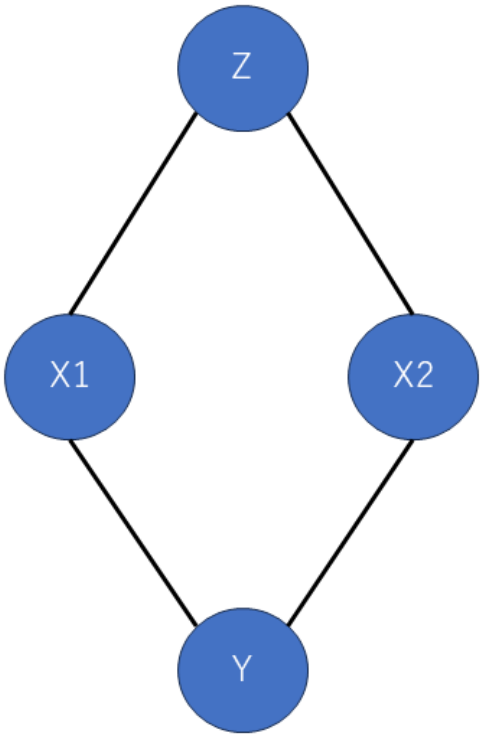

図 1: マルコフネットワークの真の構造

タを設定した. 反復回数の停止点は、Kolmogorov–Smirnov 検定統計量の平均を最小化する点に設定した. ニューラルネットワークは [9] の論文を参考にした. Python で実装し, 隠れ層 1 層 10 ノード, epochs100, batch32 に設定した.

また, Intel® Core™ i9-9900KS CPU @ 4.00GHz の 8 コアプロセッサと約 65GB のメモリを搭載した PC で実験を行った.

5.1 マルコフネットワークを用いたシミュレーション実験

この実験では, 図 1 に示すマルコフネットワークを真のモデルとした実験を行う. 図 1 のマルコフネットワークはベイジアンネットワークでは表現できない条件付き独立性を表現する. したがって, 真のモデルをすべての手法と異なるものにできる. マルコフネットワークは無向グラフで表され, ノードは確率変数を, エッジはそれらの相互依存関係を示す. マルコフネットワークはノード集合 $\mathbf{V}$ とエッジ集合 $\mathbf{E}$ を持つ無向グラフ $G = (\mathbf{V}, \mathbf{E})$ であり, 各クリーク (最大の完全部分グラフ) C について、確率変数の集合 $\mathbf{X}_C$ に対応するファクター $\phi_C(\mathbf{X}_C)$ を持つ. ネットワーク全体の同時確率分布は次のよ

表 1: 2 値の状態数を持つマルコフネットワークに非一様性の高い分布を設定した場合における各クリークの同時確率

X1	Z	P(X1, Z)
0	0	0.401416273
1	0	0.124965287
0	1	0.182032769
1	1	0.291585671

(a) X1 と Z のクリーク

Z	X2	P(Z, X2)
0	0	0.170161066
1	0	0.356220494
0	1	0.33678145
1	1	0.13683699

(b) Z と X2 のクリーク

X1	Y	P(X1, Y)
0	0	0.324909747
1	0	0.22160511
0	1	0.258539295
1	1	0.194945848

(c) X1 と Y のクリーク

Y	X2	P(Y, X2)
0	0	0.373229658
1	0	0.133712858
0	1	0.173285199
1	1	0.319772285

(d) Y と X2 のクリーク

うに表される.

$$P(X_1, X_2, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(\mathbf{X}_C) \tag{16}$$

ここで

$$Z = \sum_{X_1, X_2, \ldots, X_n} \prod_{C \in \mathcal{C}} \phi_C(\mathbf{X}_C)$$

$\mathcal{C}$ はグラフのすべてのクリークの集合である. $\phi_C(X_C)$ により変数間の依存関係が表現される.

本実験ではマルコフネットワークの同時確率が一様な場合と, 非一様な場合かつ変数の取る値が 2 値と 3 値の場合で実験を行う. マルコフネットワークの同時確率が一様な場合はすべてのパターンで同時確率が同じになる. 非一様で変数の取る値が 2 値の場合は各クリークについて表 1 のような分布になる. 非一様で変数の取る値が 3 値の場合は表 2 のような分布になる.

実験手順は以下のとおりである.

1. 図 1 のネットワークから対象の数が 100, 1000, 10000, 100000 の各場合で訓練, テスト用のデータセットを発生させる.

表 2: 3 値の状態数を持つマルコフネットワークに非一様性の高い分布を設定した場合における各クリークの同時確率

X1	Z	P(X1, Z)
0	0	0.3367290142
1	0	0.2853693180
2	0	0.0425847227
0	1	0.1107941054
1	1	0.0863571612
2	1	0.1381656784

(a) X1 と Z のクリーク

X2	Z	P(Z, X2)
0	0	0.0238610751
1	0	0.3795427874
2	0	0.2612791925
0	1	0.1452594819
1	1	0.1009957589
2	1	0.0890617043

(b) X2 と Z のクリーク

X2	Y	P(X2, Y)
0	0	0.0039066315
1	0	0.2591448623
2	0	0.2351296228
0	1	0.1254042164
1	1	0.0326806969
2	1	0.0020618193
0	2	0.0398097091
1	2	0.1887129870
2	2	0.1131494546

(c) X2 と Y のクリーク

X1	Y	P(Y, X1)
0	0	0.3365188128
1	0	0.1516498277
2	0	0.0100124761
0	1	0.0222276367
1	1	0.0336715563
2	1	0.1042475397
0	2	0.0887766701
1	2	0.1864050952
2	2	0.0664903853

(d) X1 と Y のクリーク

2. 各手法を用いて訓練データでモデルを学習し, テストデータに適用することで傾向スコアを推定する.
3. 2 で推定した傾向スコアを用いて ATE を IPW で推定する.
4. 1〜3 を 10 回繰り返し, 推定した ATE に対し, Bias, RMSE, MAE で評価した.

ここで Bias の定義は以下の式で与えられる.

$$Bias = E[\hat{ATE} - ATE_{true}] \tag{17}$$

ここで $\hat{ATE}$ は推定された ATE, ATE_{true} は真の ATE である.

表 3: マルコフネットワークに一様性の高い分布を設定した場合における各手法の推定した ATE に対する Bias, RMSE, MAE

samplesize	LR	NN	BOOST	GBN	提案手法
			Bias		
100	0.0990	0.1014	0.1009	**0.0898**	**0.0898**
1000	-0.0087	-0.0101	**-0.0082**	-0.0084	-0.0084
10000	0.0039	**-0.0003**	0.0038	0.0039	0.0039
100000	**-0.0014**	-0.0059	MO	**-0.0014**	**-0.0014**
			RMSE		
100	0.2297	0.2311	0.2367	**0.2209**	**0.2209**
1000	0.0426	0.0426	**0.0421**	0.0423	0.0423
10000	**0.0188**	0.0267	**0.0188**	**0.0188**	**0.0188**
100000	**0.0044**	0.0094	MO	**0.0044**	**0.0044**
			MAE		
100	0.1769	0.1880	0.1807	**0.1663**	**0.1663**
1000	0.0319	0.0318	0.0311	**0.0309**	**0.0309**
10000	0.0157	0.0229	**0.0156**	0.0157	0.0157
100000	**0.0038**	0.0079	MO	**0.0038**	**0.0038**

5.1.1 マルコフネットワークに一様性の高い分布を設定した場合

実験の結果を表 3 に示す. 表中の MO はメモリオーバーにより学習ができなかったことを表す.

表 3 より, サンプルサイズが 100 の場合は, GBN と提案手法が全ての指標で最も高精度であることがわかる. サンプルサイズが大きくなるとすべての指標において LR, BOOST, GBN, 提案手法の差はほとんどなくなる.

次に, 提案手法が傾向スコアを高精度に推定できることを確認するために, 推定した傾向スコアに対し, カルバックライブラー (KL) 情報量を推定した. その結果を表 4 に示す.

表 4 より, すべてのサンプルサイズにおいて, GBN と提案手法の KL 情報量は比較手法の KL 情報量以下であることが確認できる. サンプルサイズ 10000 以上はすべての手法に

表 4: マルコフネットワークに一様性の高い分布を設定した場合における各手法の推定した傾向スコアに対する KL 情報量

samplesize	LR	NN	BOOST	GBN	提案手法
100	0.0148	0.0133	0.0247	**0.0081**	**0.0081**
1000	0.0014	0.0017	0.0017	**0.0003**	**0.0003**
10000	**0.0001**	0.0003	0.0002	**0.0001**	**0.0001**
100000	**0.0000**	**0.0000**	MO	**0.0000**	**0.0000**

おいて差はほとんどない. 以上のことから提案手法はサンプルサイズが小さいときから真の傾向スコアを高精度に推定でき, それによって ATE も高精度に推定できると考えられる. また, 真のモデルが一様の場合はサンプルサイズが大きいと手法によらず高精度に傾向スコアと ATE を推定できる.

5.1.2 マルコフネットワークに非一様性（偏り）の高い分布を設定した場合 (2 値)

実験の結果を表 5 に示す.

表 5 より, サンプルサイズが 1000 のときの RMSE と MAE, 10000 以上のときの全ての指標で提案手法が最も高精度であることがわかる. この結果は前節同様, 提案手法がサンプルサイズが十分に大きいときに最も高精度に傾向スコアを推定できるためであると考えられる. ただ, サンプルサイズが 100 のときは全ての指標で NN の精度が高く, サンプルサイズが 1000 の時は Bias で GBN の精度が高いことが確認できる.

次に, 提案手法が傾向スコアを高精度に推定できることを確認するために, 推定した傾向スコアに対し, KL 情報量を推定した. その結果を表 6 に示す.

表 6 より, サンプルサイズが 1000 と 100000 の場合は提案手法の KL 情報量は最も精度が高いことが確認できる. この結果は, 提案手法が真のモデルが BN でない場合でも漸近的に真の傾向スコアを推定できるためである.

また, 表 6 より, サンプルサイズが 10000 の場合は, 提案手法の傾向スコアよりも LR の傾向スコアの方が KL 情報が小さいことがわかる. しかし, 表 5 より, サンプルサイズ 10000 における提案手法の ATE 推定の精度は LR のそれより高い. 以下ではこの原因を考察する.

(8) 式より傾向スコアを 0 または 1 に近い極端な値に推定すると, 同じ絶対誤差を持つ 0.5 に近い推定値を用いる場合に比べて, 個別介入効果 (ITE) の絶対値が過剰に大きく

表 5: 2 値の状態数をもつマルコフネットワークに非一様性の高い分布を設定した場合における各手法の推定した ATE に対する Bias, RMSE, MAE

samplesize	LR	NN	BOOST	GBN	提案手法
			Bias		
100	0.0227	**0.0189**	0.0983	-0.6544	-0.6544
1000	-0.0134	-0.0076	-0.0056	**-0.0018**	-0.0163
10000	0.0029	0.0025	0.0053	0.0198	**-0.0022**
100000	-0.0003	-0.0010	MO	0.0016	**0.0000**
			RMSE		
100	0.1774	**0.1646**	0.2592	2.2746	2.2746
1000	0.0671	0.0693	0.0695	0.0737	**0.0664**
10000	0.0186	0.0195	0.0189	0.0286	**0.0182**
100000	0.0054	0.0144	MO	0.0073	**0.0049**
			MAE		
100	0.1447	**0.1355**	0.1896	0.8669	0.8669
1000	0.0592	0.0592	0.0605	0.0641	**0.0581**
10000	0.0159	0.0162	0.0169	0.0245	**0.0151**
100000	0.0044	0.0125	MO	0.0064	**0.0043**

なってしまう場合がある．ここで ITE とは (8) 式の ATE の平均をとらないもので, 各データの介入効果である．ITE は以下の式で表せる．

$$ITE = \frac{y_1}{e(x)}z - \frac{y_0}{1-e(x)}(1-z) \tag{18}$$

以下では, 真の傾向スコアに比べて 0.5 に近い値に推定している (inner) 場合と 0 または 1 に近い値に推定している (outer) 場合の個数と RMSE を推定した．その結果を表 7, 表 8 にそれぞれ示す．

サンプルサイズ 10000 のとき, 表 7 より, 提案手法の outer の個数が最も少なく, 表 8 より, 提案手法の outer の RMSE の精度が最も高いことがわかる．LR は傾向スコアの推定値が真の傾向スコアより外側に大きく推定しているものが多いため, ATE 推定精度が悪くなると考えられる．

表 6: 2 値の状態数をもつマルコフネットワークに非一様性の高い分布を設定した場合における各手法の推定した傾向スコアに対する KL 情報量

samplesize	LR	NN	BOOST	GBN	提案手法
100	**0.0164**	0.0243	0.0178	0.0532	0.0532
1000	0.0013	0.0016	0.0016	0.0019	**0.0012**
10000	**0.0001**	0.0004	0.0002	0.0007	0.0003
100000	**0.0000**	0.0001	MO	**0.0000**	**0.0000**

表 7: 2 値の状態数をもつマルコフネットワークに非一様性の高い分布を設定した場合における個数（inner および outer）に基づく各手法の比較

個数 inner					
samplesize	LR	NN	BOOST	GBN	NCPMIN
100	22.4	27.1	**13.0**	17.3	17.3
1000	156.4	144.8	**124.5**	157.7	177.9
10000	1740.1	1149.8	**1,021.0**	1305.8	2294.4
100000	9554.5	15664.6	MO	**8850.5**	10818.5
個数 outer					
samplesize	LR	NN	BOOST	GBN	NCPMIN
100	4.7	**0.0**	14.1	9.8	9.8
1000	106.4	118.0	138.3	105.1	**84.9**
10000	950.2	1540.5	1,669.3	1384.5	**395.9**
100000	17450.5	**11340.4**	MO	18154.5	16186.5

5.1.3 マルコフネットワークに非一様性（偏り）の高い分布を設定した場合 (3 値)

実験の結果を表 9 に示す.

表 9 より, サンプルサイズが 10000 以上の場合は, 提案手法が全ての指標で最も高精度であることがわかる. この結果は, 提案手法がサンプルサイズが十分に大きいときに最も高精度に傾向スコアを推定できるためであると考えられる. ただ, サンプルサイズが 100

表 8: 2 値の状態数をもつマルコフネットワークに非一様性の高い分布を設定した場合における RMSE（inner および outer）に基づく各手法の比較

RMSE inner					
samplesize	LR	NN	BOOST	GBN	NCPMIN
100	0.0722	0.0931	**0.0381**	0.0665	0.0665
1000	0.0114	0.0083	**0.0075**	0.0128	0.0131
10000	**0.0028**	0.0049	0.0051	0.0116	0.0080
100000	**0.0007**	0.0047	MO	0.0019	**0.0007**
RMSE outer					
samplesize	LR	NN	BOOST	GBN	NCPMIN
100	0.0039	**0.0000**	0.0332	0.0296	0.0296
1000	0.0071	0.0113	0.0115	0.0089	**0.0033**
10000	0.0027	0.0049	0.0041	0.0066	**0.0002**
100000	**0.0012**	0.0031	MO	0.0018	0.0014

のときは全ての指標で NN の精度が高く, サンプルサイズが 1000 の時は RMSE, MAE で LR の精度が高いことが確認できる.

次に, 提案手法が傾向スコアを高精度に推定できることを確認するために, 推定した傾向スコアに対し, カルバックライブラー (KL) 情報量を推定した. その結果を表 10 に示す.

表 10 より, サンプルサイズが 10000 以上の場合は提案手法の KL 情報量は比較手法の KL 情報量以下であることが確認できる. この結果は, 提案手法が真のモデルが BN でない場合でも漸近的に真の傾向スコアを推定できるためである.

また, 表 10 より, サンプルサイズが 100, 1000 の場合には BOOST の傾向スコアの推定精度が最も高いことがわかる. しかし, 表 9 よりサンプルサイズ 100, 1000 における BOOST の ATE 推定の精度は提案手法に比べて低い. これは, BOOST が傾向スコアを 0 や 1 に極端に近い値に推定する場合があり, (8) 式の傾向スコアによる逆確率重み付けが極端に大きくなる場合があるからだと考えられる.

次に, BOOST が傾向スコアを 0 や 1 に近い値に推定することで ATE の精度を低下させていることを確認するために, サンプルサイズ 100 のときの各手法における傾向スコアと個別介入効果 (ITE) の分布を推定した. その結果を図 2 に示す.

表 9: 3 値の状態数をもつマルコフネットワークに非一様性の高い分布を設定した場合における各手法の推定した ATE に対する Bias, RMSE, MAE

samplesize	LR	NN	BOOST	GBN	提案手法
			Bias		
100	0.3297	**0.1844**	3.6938	-0.4504	-0.5406
1000	0.0652	0.0352	-0.0403	0.0266	**-0.0039**
10000	0.0494	0.0288	0.0216	0.0385	**0.0211**
100000	0.0342	0.0194	MO	0.0032	**0.0008**
			RMSE		
100	0.5378	**0.3955**	11.1671	1.7477	1.7484
1000	**0.1350**	0.1430	0.2831	0.3318	0.3183
10000	0.0588	0.0722	0.0401	0.0669	**0.0400**
100000	0.0352	0.0374	MO	0.0092	**0.0089**
			MAE		
100	0.4285	**0.3254**	4.0066	0.8779	0.8805
1000	**0.0963**	0.1271	0.1738	0.2240	0.2112
10000	0.0494	0.0566	0.0337	0.0529	**0.0336**
100000	0.0342	0.0322	MO	0.0076	**0.0068**

図 2 より, BOOST は傾向スコアを 0 に近い値に推定し, ITE が極端に大きくなる場合があることが確認できる. この極端に大きい ITE により BOOST の ATE 推定精度が低くなったのだと考えられる.

マルコフネットワークが一様分布の場合, サンプルサイズが小さいときからほかの手法より ATE 推定精度が高く, サンプルサイズが大きくなるとほかの手法と差が小さくなる. マルコフネットワークが非一様分布の場合, サンプルサイズが大きいと ATE 推定精度が高くなる. ここから, 提案手法は真のモデルが単純なときはサンプルサイズが小さくても真の傾向スコアを高精度に推定でき, 真のモデルが複雑な時は漸近的に真の傾向スコアに一致するといえる.

表 10: 3 値の状態数をもつマルコフネットワークに非一様性の高い分布を設定した場合における各手法の推定した傾向スコアに対する KL 情報量

samplesize	LR	NN	BOOST	GBN	提案手法
100	0.1101	0.1099	**0.0561**	0.1382	0.2107
1000	0.0924	0.0540	**0.0064**	0.0273	0.0308
10000	0.0890	0.0101	**0.0006**	0.0053	**0.0006**
100000	0.0885	0.0055	MO	0.0004	**0.0001**

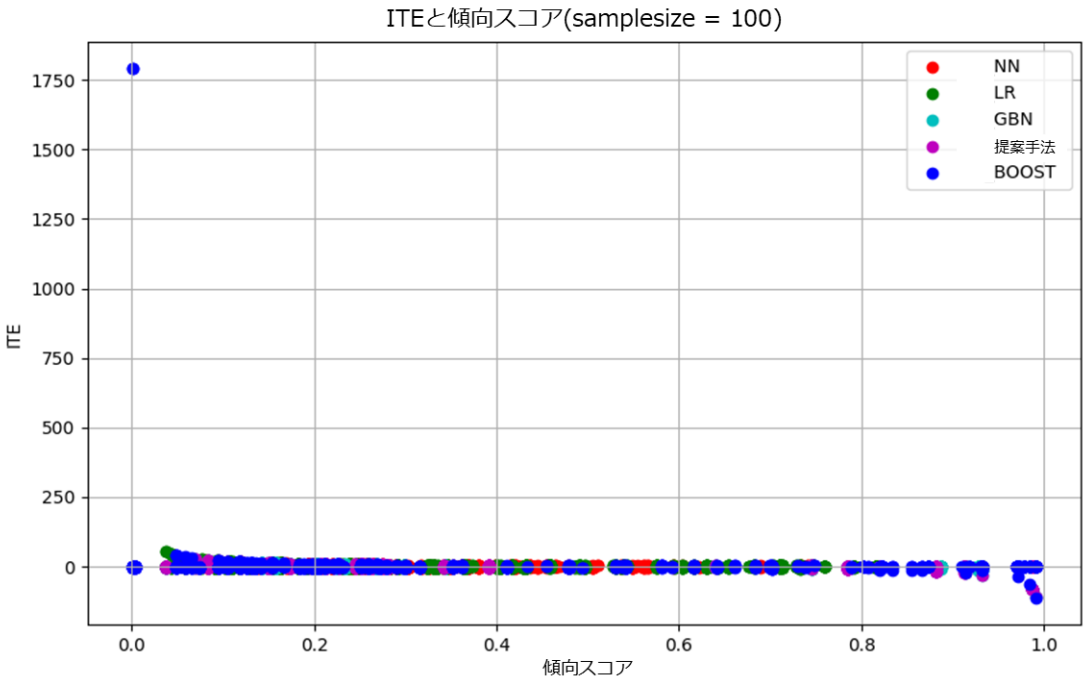

図 2: サンプルサイズ 100 の各手法の傾向スコアと個別介入効果 (ITE)

5.2 実データ jobs を用いた評価実験

次に, 実データでの有効性を評価するために, 因果推論分野でよく用いられる jobs データセットを用いて評価を行う. jobs データセット [3, 24, 25] は, 職業訓練プログラムの介入効果を評価するために使用され, 年齢, 教育レベル, 収入, 婚姻状況などの共変量が含まれている. 介入は職業訓練への参加の有無, 効果は介入後に就職できたかどうかである.

表 11: 実データ jobs における各手法の推定した ATE に対する Bias, RMSE, MAE

	LR	NN	BOOST	GBN	提案手法
Bias	-0.3755	-0.5486	-0.2939	13.0158	**0.0060**
RMSE	0.3883	0.5530	0.3384	26.5959	**0.1877**
MAE	0.3755	0.5486	0.2939	13.4092	**0.1515**

jobs データセットでは, 無作為化比較試験が実施された介入群 297 件と対照群 425 件から真の ATE を求め, このデータセットに無作為化比較試験ではない状況で実施された 2490 件の対照群を加えることで, バイアスのあるデータセットを作成している. 真の ATE は 0.0779 であり, 職業訓練の効果が認められるが, jobs データセット全体に対し, 介入群の結果の期待値から対照群の結果の期待値を単純に計算すると, -0.0832 となり職業訓練をすることで負の効果が出るという誤った結論となる. 本実験では, このデータセットを訓練データとテストデータに 8:2 の割合で分割し, 前節と同様の手順, 評価指標で ATE の評価を行った.

ATE の推定結果を表 11 に示す. 表 11 より, 提案手法の ATE が Bias, RMSE, MAE の全ての指標で最も精度が高かった. これは, 提案手法が漸近的に真の傾向スコアを推定できることにより, 比較手法に比べて傾向スコアを高精度に推定できたためだと考えられる. また, 自動的な変数選択により傾向スコア推定に関係のある変数のみを用いて推定できたことも要因として考えられる.

また, Bias の定義より,

$$E[\hat{ATE}] = Bias + ATE_{true} \tag{19}$$

が成り立つ. jobs データセットの真の ATE は 0.0779 であるので, 式 (19) と表 11 における Bias の結果から, 職業訓練に正の効果があると推定できたのは GBN と提案手法のみであることが確認できる. LR, NN, BOOST についてはいずれも職業訓練を行うことで就職に負の効果が出ていると推定しており, バイアスを除けていないことがわかる.

次に, 提案手法で学習した構造を確認し, 推定に用いた変数を確認する. 学習した構造を図 3 に示す. 図 3 における z は介入の有無を表す二値変数, y は介入後に就職できたかの二値変数である. また, $black$ は黒人の方か, $married$ は結婚をしているか, $hispanic$ はヒスパニックの方か, age は年齢, $nodegree$ は学位を持っているかの二値変数である. また, age, $education$, $re75$ はそれぞれ年齢, 教育のレベル, 1975 年における収入を表す変数である.

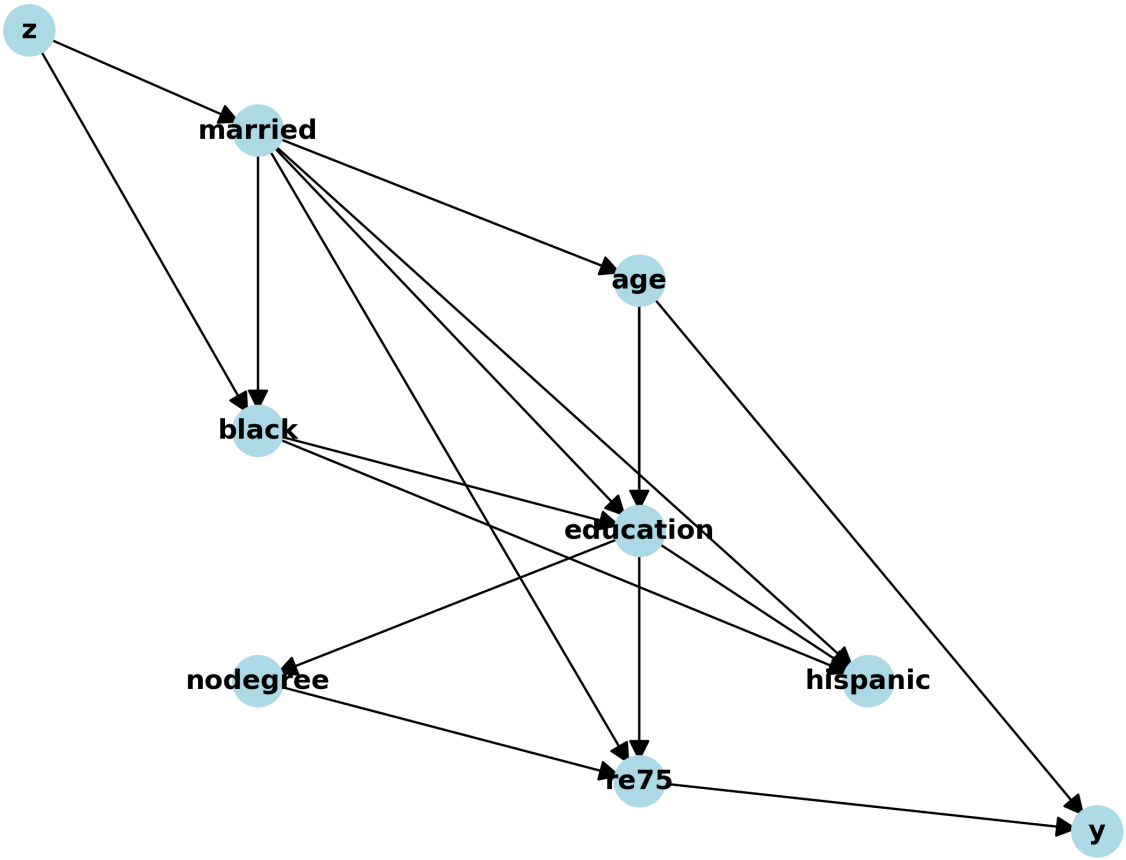

図 3: jobs データで学習した構造

ベイジアンネットワーク分類器では目的変数のマルコフブランケットに属する変数を用いて推定を行う．したがって, 図 3 より, 提案手法は対象が黒人の方か, 結婚をしているかどうかで傾向スコアの推定を行ったことがわかった．また, z と y が共変量を所与として条件付き独立になっていることも確認できた.

5.3 実データ twins を用いた評価実験

次に, 因果推論分野でよく用いられる twins データセットを用いて評価を行う．twins データセットは, 1989 年から 1991 年の期間における米国での全出生記録から得られたものである [30]．乳幼児の出生体重による死亡率を評価するために使用され, 年齢, 性別, 母親の健康状態, 妊娠, 両親の社会経済的背景などの共変量が含まれている．

まず, 出生体重が 2kg 未満の同性の双子を選択し, 各双子について両親, 妊娠, 出生に関する 46 の共変量を取得した．欠損値のあるデータを除外することで完全なデータセット

表 12: 実データ twins における各手法の推定した ATE に対する Bias, RMSE, MAE

	LR	NN	BOOST	提案手法
Bias	0.0689	0.0498	0.0317	**0.0051**
RMSE	0.0784	0.0984	0.0439	**0.0417**
MAE	0.0689	0.0796	0.0393	**0.0328**

を作成した. 最終的なデータセットは 4821 組の双子ペアで構成されており, 軽い方の双子の死亡率は 16.91 %, 重い方は 14.42 % である. このデータセットでは, 各双子ペアについて軽い方 ($z = 0$) と重い方 ($z = 1$) の両方の結果が観測されているため, ATE の真値が既知である. 真の ATE は-0.0248 であり, 双子のうち体重が軽い方が死亡率が高いという結果になる.

観察研究をシミュレーションするために共変量を用いて選択バイアスを導入した. 具体的には, 各双子ペアについて以下の式 (20) に基づいて軽い方 ($z = 0$) または重い方 ($z = 1$) のどちらか一方を観測する [28]. 共変量ベクトルを $\mathbf{x}$, 共変量の個数を d とすると,

$$z \mid \mathbf{x} \sim \mathrm{Bern}(\mathrm{Sigmoid}(\mathbf{w}^\top \mathbf{x} + n)) \tag{20}$$

ここで,

$$\mathbf{w}^\top \sim \mathcal{U}((0.1, 0.1)^{d\times 1}), \quad n \sim \mathcal{N}(0, 0.1)$$

以上の方法により得られたデータセットを用いて, 提案手法および比較手法の評価を行った. 本実験では, このデータセットを訓練データとテストデータに 8:2 の割合で分割し, 前節と同様の手順, 評価指標で ATE の評価を行った.

ATE の推定結果を表 12 に示す. 表 12 より, 提案手法の ATE が Bias, RMSE, MAE の全ての指標で最も精度が高かった. また, twins データセットの真の ATE は-0.0248 であるので, 式 (19) と表 12 における Bias の結果から, 体重が重い方が死亡率が低いと推定できたのは提案手法のみであることが確認できる. LR, NN, BOOST についてはいずれも体重が重い方が死亡率が高いと推定している.

次に, 提案手法で学習した構造を確認する. 学習した構造を図 4 に示す. 変数数が 46 個と多いため, 目的変数のマルコフブランケットに属する変数のみの構造を図 5 に示す. 図 5 における z は介入の有無を表す二値変数, y は介入後に死亡したかの二値変数である. また, $lung$ は急性および慢性の肺疾患があるか, $renal$ は腎臓病かどうか, $alcohol$ は飲酒

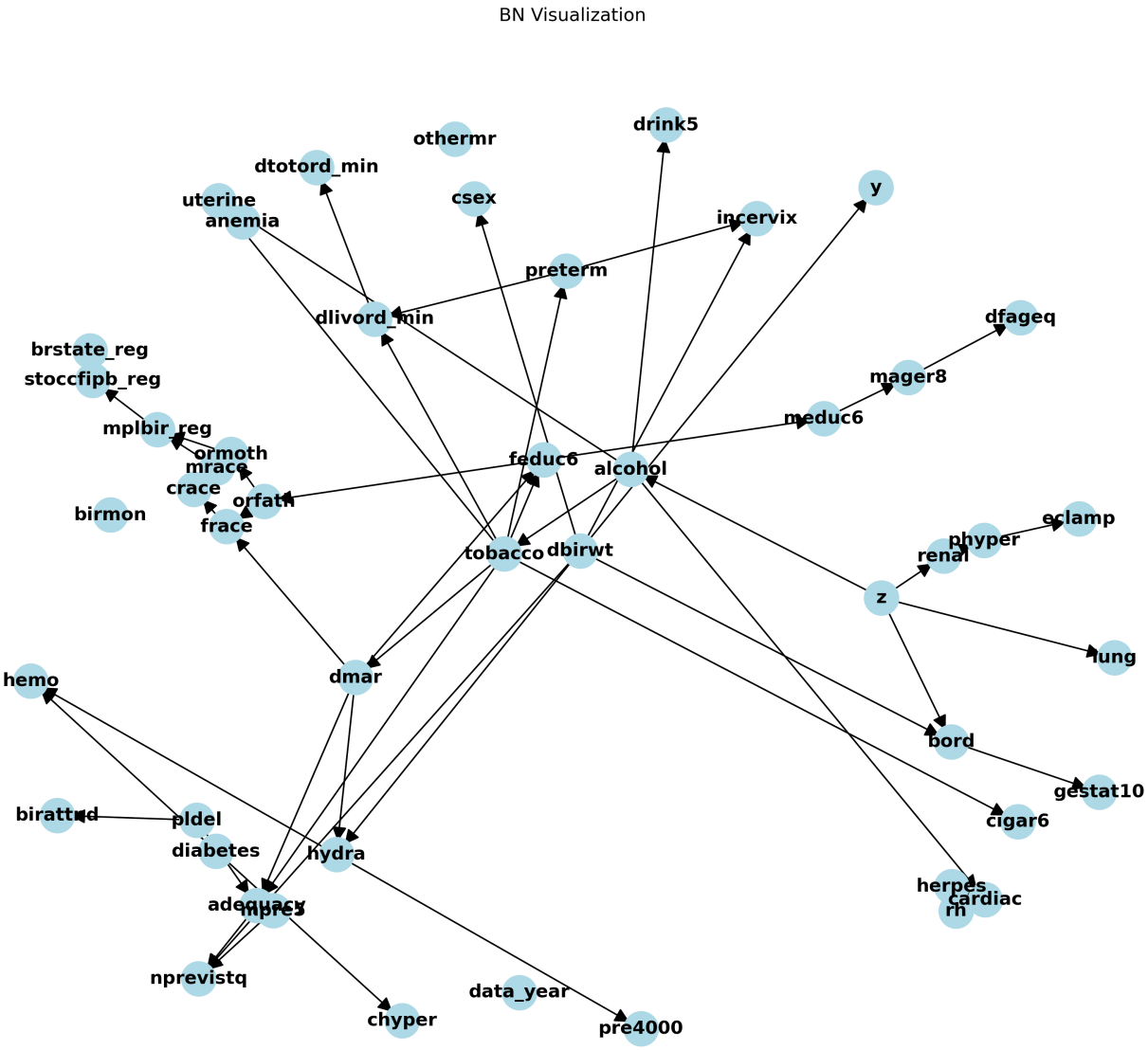

図 4: twins データで学習した構造

したか, *bord* は生まれた順番, *dbirwt* は子の体重である. 提案手法はこれらの共変量で推定を行ったことが学習したグラフ構造からわかる. また, z と y が共変量を所与として条件付き独立になっていることも確認できた.

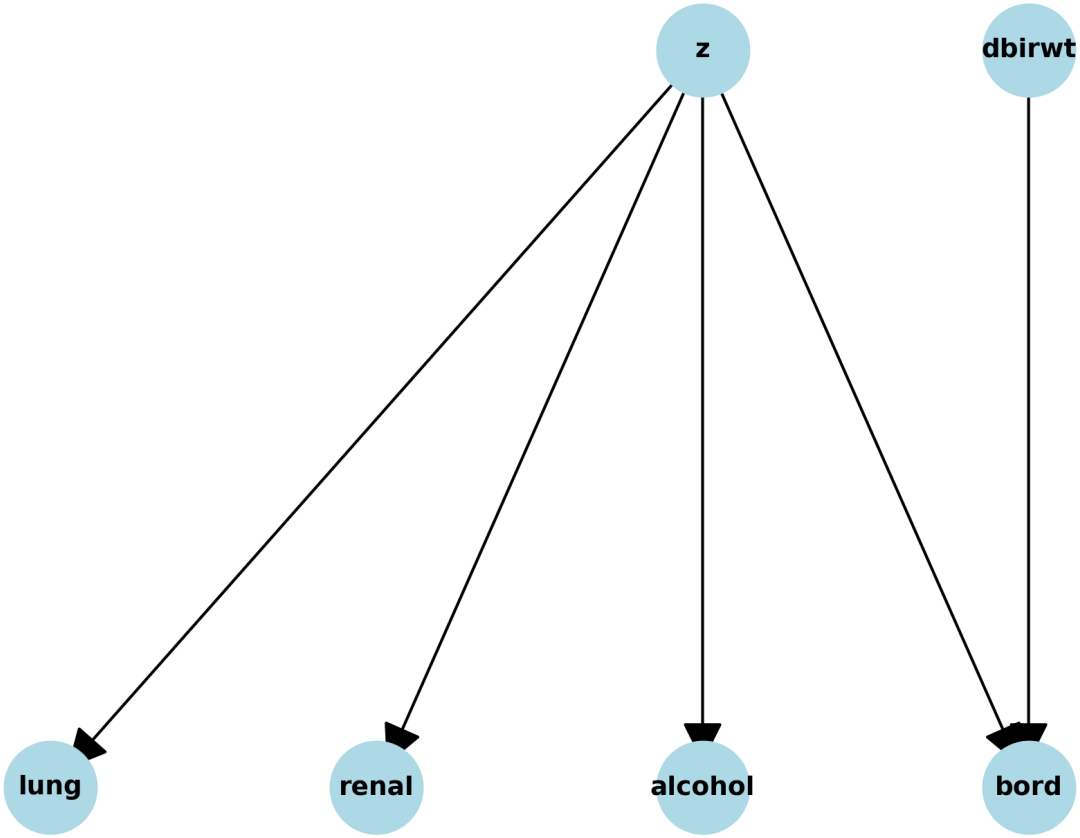

図 5: twins データの目的変数のマルコフブランケット

6　むすび

本研究では, 因果推論における傾向スコア推定の精度向上を目的として, 真の傾向スコアに漸近的に一致するベイジアンネットワーク分類器 (BNC) を用いた新しい手法を提案した.

提案手法は以下の利点をもつ.

1. 真のモデルがベイジアンネットワークに従っていなくても, 漸近的に真の傾向スコアを得ることが可能である.
2. 構造学習による自動的な変数選択が可能である.

シミュレーション実験と jobs データセット, twins データセットを用いた実験の結果, 提案手法が他の機械学習手法に比べて一貫して高い ATE 推定精度を示した. 今後の研究で

は, 提案手法を連続値データに対応させることや, さらに複雑な実社会データセットに適用し, その汎用性を検証することを目指す.

参考文献

[1] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[2] Patricia A Gum, Maran Thamilarasan, Junko Watanabe, Eugene H Blackstone, and Michael S Lauer. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. *Jama*, 286(10):1187–1194, 2001.

[3] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[4] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[5] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[6] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[7] Pedro HC Sant'Anna, Xiaojun Song, and Qi Xu. Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6):1093–1120, 2022.

[8] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[9] Soko Setoguchi, Sebastian Schneeweiss, M Alan Brookhart, Robert J Glynn, and E Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555, 2008.

[10] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score

estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8):826, 2010.

[11] Shouta Sugahara, Koya Kato, and Maomi Ueno. Learning bayesian network classifiers to minimize the class variable parameters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20540–20549, 2024.

[12] John Z Ayanian, Mary Beth Landrum, Edward Guadagnoli, and Peter Gaccione. Specialty of ambulatory care physicians and mortality among elderly patients after myocardial infarction. *New England journal of medicine*, 347(21):1678–1686, 2002.

[13] Jyotsna Jalan and Martin Ravallion. Estimating the benefit incidence of an antipoverty program by propensity-score matching. *Journal of Business & Economic Statistics*, 21(1):19–30, 2003.

[14] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[15] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[16] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[17] Insung Kong, Yuha Park, Joonhyuk Jung, Kwonsang Lee, and Yongdai Kim. Covariate balancing using the integral probability metric for causal inference. In *International Conference on Machine Learning*, pages 17430–17461. PMLR, 2023.

[18] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[19] Christiana Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, pages 1231–1236, 1993.

[20] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[21] Shouta Sugahara and Maomi Ueno. Exact learning augmented naive bayes classifier. *Entropy*, 23(12):1703, 2021.

[22] Shouta Sugahara, Wakaba Kishida, Koya Kato, and Maomi Ueno. Recursive au-

tonomy identification-based learning of augmented naive bayes classifiers. In *International Conference on Probabilistic Graphical Models*, pages 265–276. PMLR, 2022.

[23] W. Buntine. Theory refinement on Bayesian networks. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, pages 52–60, 1991.

[24] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for non-experimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002.

[25] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

[26] Uri Shalit, Fredrik Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3076–3085, 2017.

[27] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6446–6456, 2017.

[28] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, 2018.

[29] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

[30] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.