二値データに対する因果推論手法の提案

稲積 孝紀[*1], 鷲尾 隆[*1], 清水 昌平[*1], 鈴木 譲[*2], 山本 章博[*3], 河原 吉伸[*1]

[*1] 大阪大学 産業科学研究所

[*2] 大阪大学 大学院理学研究科

[*3] 京都大学 大学院情報学研究科

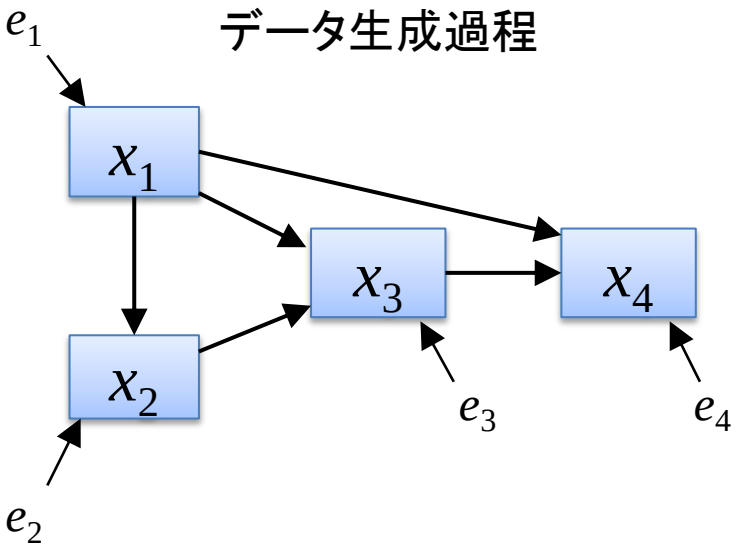
研究背景と目的

- 与えられた二値データ $X = \{x_1, x_2, \ldots, x_d\}$ が
  どのようなデータ生成過程によって生成されているかを
  推定する.

二値データ

x_1	x_2	x_3	x_4
0	1	0	0
1	0	1	0
1	0	0	1
1	1	1	0
⋮	⋮	⋮	⋮

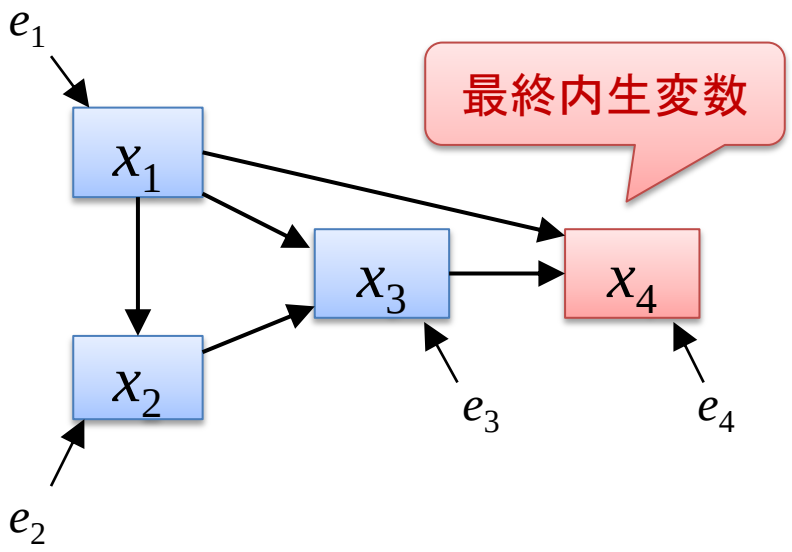
推定



- 従来手法 … edge の向きを一意に同定できない場合がある.
  → データ生成過程を一意に同定できる
  新しい因果推論手法 BExSAM を提案する（UAI2011 採択）.

提案手法: BExSAM モデル

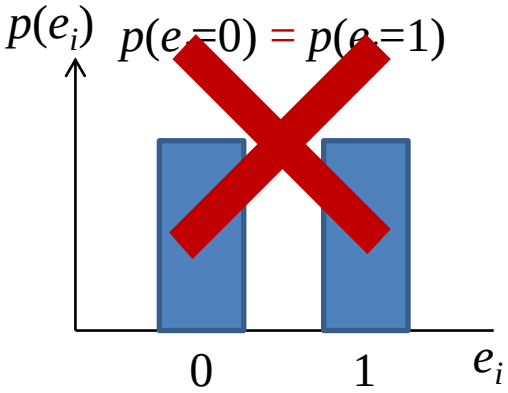
- 因果構造として DAG 構造を考え，対応する式を定義する.
  - 外乱 e_i は排他的論理和によって加わるものとする.



対応

$$x_1 = e_1$$
$$x_2 = x_1 \oplus e_2$$
$$x_3 = x_1\, x_2 \oplus e_3$$
$$x_4 = (x_1 + x_3) \oplus e_4$$

- さらに，外乱 e_i の分布に偏りを仮定する:
  $0 < p(e_i=1) < 1$ かつ $p(e_i=1) \neq 0.5$



提案手法: モデル推定アルゴリズム

- 最終内生変数 x_k の同定

x_2 の比が $x_1 = 0, 1$ について
項の交換を許した上で等しい.

x_2 は最終内生変数である.

観測データの度数表

x_1	x_2	出現回数	x_2 の比
0	0	20	2 : 1
0	1	10	
1	0	15	1 : 2
1	1	30	

- 最終内生変数 x_k の親となる変数集合の同定

x_2 の比が x_1 の値によって異なる.

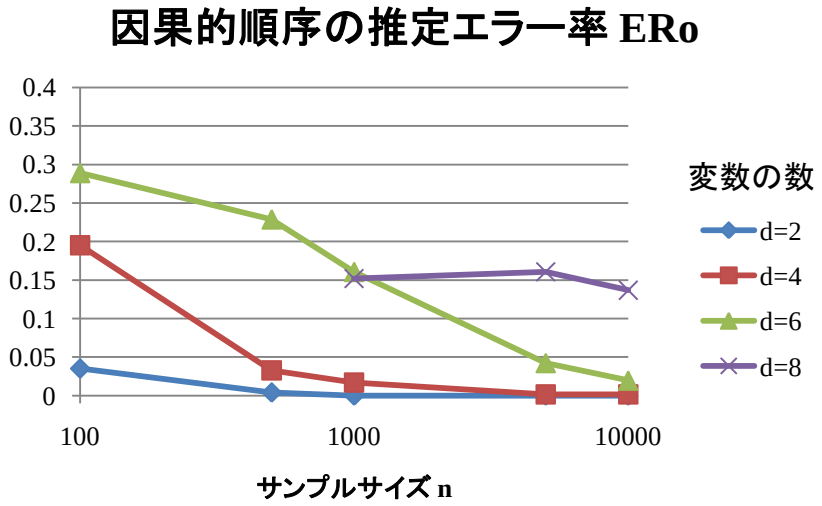
x_1 は x_2 の親である.

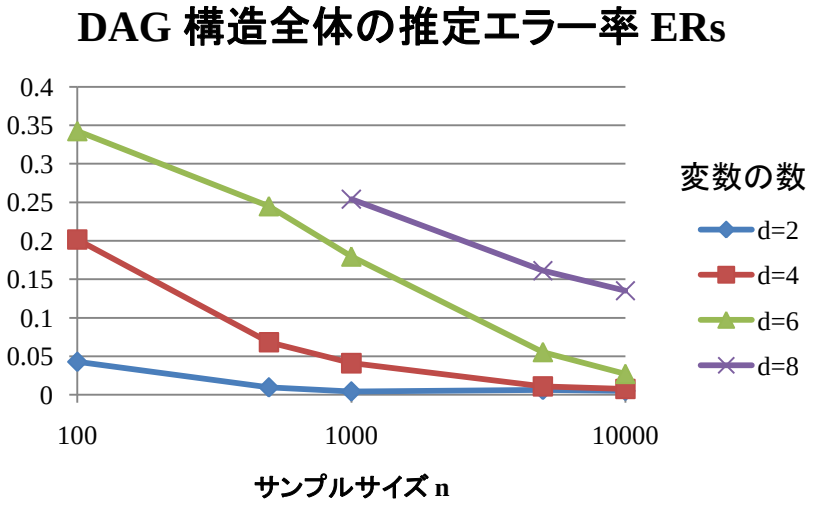
観測データの度数表

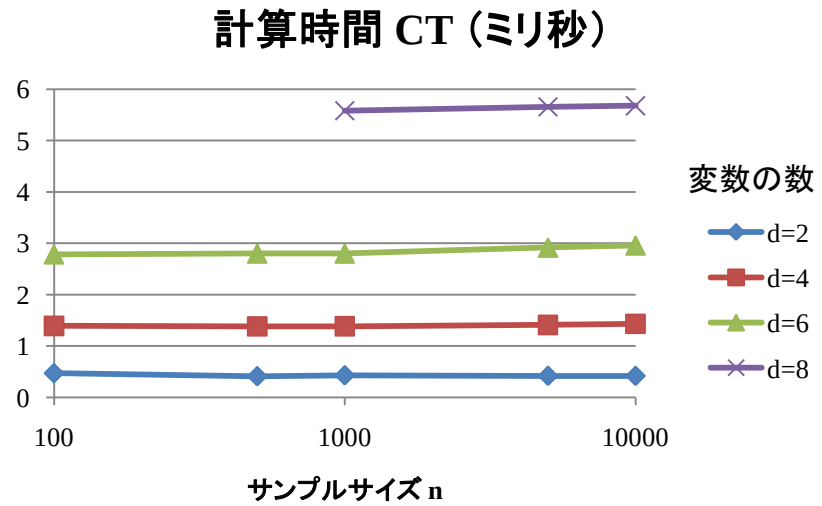
x_2	$x_1 = 0$ のときの出現回数	$x_1 = 1$ のときの出現回数
0	20	15
1	10	30

- x_k を除いたモデルを新たな入力モデルとして繰り返す.

評価実験: 人工データ







- エラー率(左上, 右上):
  - サンプルサイズの増加に従い減少する.
- 計算時間(左下):
  - 主に変数の数の増加に従い増加する.

評価実験: 実世界データ

- ウィスコンシン州の高校生の
  大学進学意思，親の働きかけ，社会的ステータスの関係
  [Sewell and Shah 1968, Spirtes et al. 2000]
  - 男子高校生のみを抽出.
  - IQ（4値変数）が最も低い学生を除く.
  - サンプルサイズ 3756
- BExSAM を適用した結果（検定の有意水準 0.05）