属性間の依存関係の推定と実装

鈴木 譲

大阪大学

2014 年 9 月 9 日
ERATO 湊離散構造処理系ワークショップ (礼文島)

目次

$P(X, Y, Z)$ の因数分解は、11 通り

X, Y, Z: (離散) 確率変数

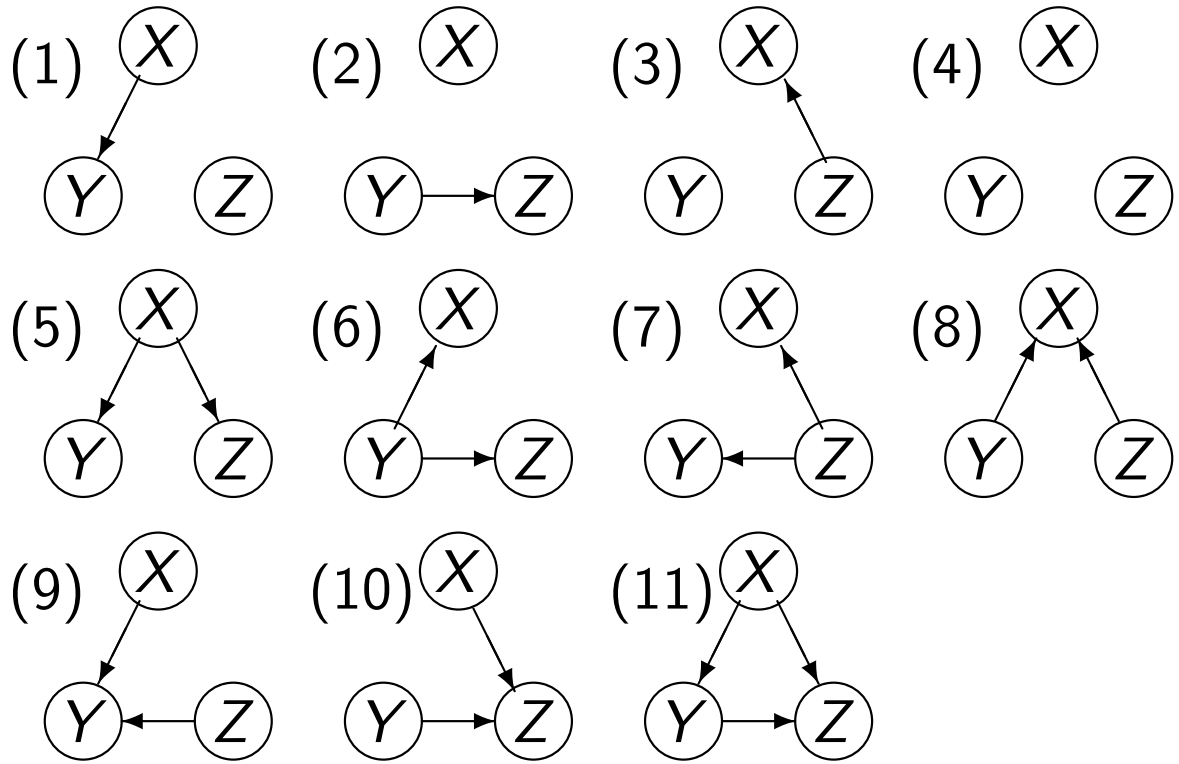
$$P(X)P(Y)P(Z), P(X)P(Y, Z), P(Y)P(Z, X),$$

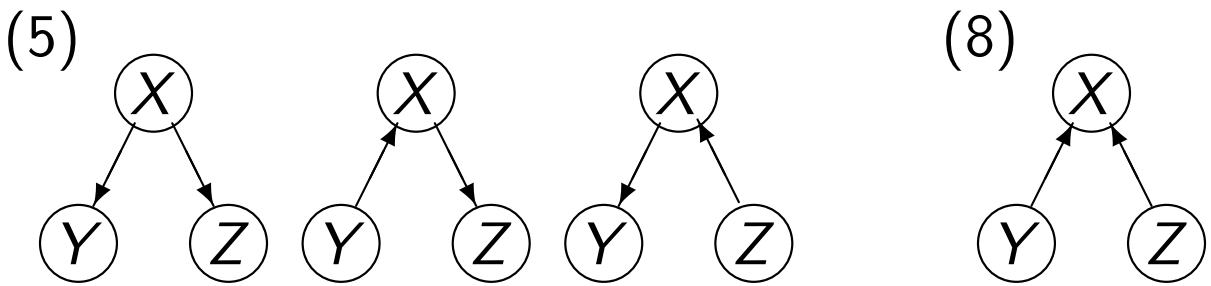
$$P(Z)P(X, Y), \frac{P(X, Y)P(X, Z)}{P(X)}, \frac{P(X, Y)P(Y, Z)}{P(Y)},$$

$$\frac{P(X, Z)P(Y, Z)}{P(Z)}, \frac{P(Y)P(Z)P(X, Y, Z)}{P(Y, Z)}, \frac{P(Z)P(X)P(X, Y, Z)}{P(Z, X)},$$

$$\frac{P(X)P(Y)P(X, Y, Z)}{P(X, Y)}, P(X, Y, Z) \ ,$$

Bayesian ネットワーク: 条件付き独立性を有向非巡回グラフで





定式化

n 組の例

$$\left.\begin{array}{lll} X = x_1 & Y = y_1 & Z = z_1 \\ X = x_2 & Y = y_2 & Z = z_2 \\ \vdots & \vdots & \vdots \\ X = x_n & Y = y_n & Z = z_n \end{array}\right\} \text{i.i.d.}$$

から、BN の構造を推定 ((1)-(11) のいずれであるかを特定)

従来の研究

- 独立性検定によるもの: PC アルゴリズム (Spirtes, 2000) 他
- Bayes によるもの
  - 各測度のスコアの計算:
    - 離散のみ: Cooper (1991), Suzuki (1993) 以降多数、ほとんどがこれ
    - Gauss のみ: 共分散を計算
    - 離散と連続が混在: Bottcher の R パッケージなど
      連続はすべて Gauss を仮定、性能の保証がない
  - 各測度のスコアの値から各構造のスコアを求め、最適な構造を見出す
    [今回の検討の対象外]

本研究: 離散や連続を区別しない BN の構造推定

従来

レコードの全部の項目が有限個の値をとる

$(X, Y, Z) =$ (性別, 昭和・大正・平成, 都道府県)
$A = \{0, 1\}$, $B = \{0, 1, 2\}$, $C = \{0, 1, \cdots, 46\}$

難しい問題から逃げている

提案

レコードの各項目が離散、連続、どちらでもないも可

$(X, Y, Z) =$ (身長, 年齢, 性別)
$A = [0, 1)$, $B = \{1, 2, \cdots\}$, $C = \{0, 1\}$

ベイズによる独立性検定

$x^n = (x_1, \cdots, x_n)$, $y^n = (y_1, \cdots, y_n)$ から、
$P(X), P(Y), P(X, Y)$ が、事前確率 $w(\cdot)$ のパラメータ θ で表現

$$Q^n(x^n) := \int P^n(x^n|\theta)w(\theta)d\theta \ , \ Q^n(y^n) := \int P^n(y^n|\theta)w(\theta)d\theta \ ,$$

$$Q^n(x^n, y^n) := \int P^n(x^n, y^n|\theta)w(\theta)d\theta \ ,$$

$X \perp\!\!\!\perp Y$ の事前確率を p として、x^n, y^n のもとでの $X \perp\!\!\!\perp Y$ の事後確率:

$$P(X \perp\!\!\!\perp Y|x^n, y^n) = \frac{pQ^n(x^n)Q(y^n)}{pQ^n(x^n)Q(y^n) + (1-p)Q^n(x^n, y^n)} \ ,$$

$X \perp\!\!\!\perp Y$ であるか否かの決定ルール:

$$X \perp\!\!\!\perp Y \Longleftrightarrow pQ^n(x^n)Q(y^n) \geq (1-p)Q^n(x^n, y^n) \ .$$

(1)-(11) の構造の同定

$p_1, \cdots, p_{11}$: 構造の事前確率
$x^n = (x_1, \cdots, x_n)$, $y^n = (y_1, \cdots, y_n)$, $z^n = (z_1, \cdots, z_n)$

$$p_1 Q^n(x^n) Q^n(y^n) Q(z^n),$$

$$p_2 Q^n(x^n) Q^n(y^n, z^n), p_3 Q^n(y^n) Q^n(z^n, x^n), p_4 Q^n(z^n) Q^n(x^n, y^n),$$

$$p_5 \frac{Q^n(x^n, y^n) Q^n(X, z^n)}{Q^n(x^n)}, p_6 \frac{Q^n(x^n, y^n) Q^n(y^n, z^n)}{Q^n(y^n)}, p_7 \frac{Q^n(x^n, z^n) Q^n(y^n, z^n)}{Q^n(z^n)},$$

$$p_8 \frac{Q^n(y^n) Q^n(z^n) Q(x^n, y^n, z^n)}{Q^n(y^n, z^n)}, p_9 \frac{Q^n(z^n) Q^n(x^n) Q^n(x^n, y^n, z^n)}{Q^n(z^n, x^n)},$$

$$p_{10} \frac{Q^n(x^n) Q^n(y^n) Q^n(x^n, y^n, z^n)}{Q^n(x^n, y^n)}, p_{11} Q^n(x^n, y^n, z^n)$$

x^n を **gzip** で圧縮したときの長さを $l(x^n)$ として、$Q = 2^{-l(x^n)}$

Kraft の不等式: $\sum_{x^n} Q^n(x^n) \leq 1$

Q^n はユニバーサル

$P(X|\theta)$ の θ がどのような値であっても、 $n \to \infty$ で確率 1 で

$$\frac{1}{n} \log \frac{P^n(x^n|\theta)}{Q^n(x^n)} \to 0$$

$n \to \infty$ で、決定ルールが
モデルの事前確率 $\{p_i\}$、パラメータの事前確率 $w(\cdot)$ によらない

X に確率密度関数 f が存在するとき (Ryabko の方法)

- $A_0 := \{A\}$
- A_{j+1} は、A_j の細分化

各レベル j で、$x^n = (x_1, \cdots, x_n)$ を $(a_1^{(j)}, \cdots, a_n^{(j)})$ に量子化

$$A_1 \quad \longrightarrow \quad g_1^n(x^n) = \frac{Q_1^n(a_1^{(1)}, \cdots, a_n^{(1)})}{\lambda(a_1^{(1)}) \cdots \lambda(a_n^{(1)})}$$

$$A_2 \quad \longrightarrow \quad g_2^n(x^n) = \frac{Q_2^n(a_1^{(2)}, \cdots, a_n^{(2)})}{\lambda(a_1^{(2)}) \cdots \lambda(a_n^{(2)})}$$

$$\vdots \qquad \vdots$$

$$A_j \quad \longrightarrow \quad g_j^n(x^n) = \frac{Q_j^n(a_1^{(j)}, \cdots, a_n^{(j)})}{\lambda(a_1^{(j)}) \cdots \lambda(a_n^{(j)})}$$

λ: Lebesgue 測度 (区間の幅)

$\sum_j w_j = 1,\ w_j > 0$ を用いて、

$$g^n(x^n) := \sum_{j=1}^{\infty} w_j g_j^n(x^n)$$

$g^n(y^n),\ g^n(z^n),\ g^n(x^n, y^n),\ g^n(x^n, z^n),\ g^n(y^n, z^n),\ g^n(x^n, y^n, z^n)$ も同様。

$$p_1 g^n(x^n) g^n(y^n) g(z^n), p_2 g^n(x^n) g^n(y^n, z^n), p_3 g^n(y^n) g^n(z^n, x^n),$$

$$p_4 g^n(z^n) g^n(x^n, y^n), p_5 \frac{g^n(x^n, y^n) g^n(X, z^n)}{g^n(x^n)}, p_6 \frac{g^n(x^n, y^n) g^n(y^n, z^n)}{g^n(y^n)},$$

$$\cdots$$

$$p_{10} \frac{g^n(x^n) g^n(y^n) g^n(x^n, y^n, z^n)}{g^n(x^n, y^n)}, p_{11} g^n(x^n, y^n, z^n)\ ,$$

予測確率密度 g^n のユニバーサル性

f: 確率密度関数
f_j (レベル j の確率密度関数)
$f^n(x^n) := f(x_1) \cdots f(x_n)$

Ryabko 2009

$D(f||f_j) \to 0$ $(j \to \infty)$ なる f について、$n \to \infty$ で確率 1 で

$$\frac{1}{n} \log \frac{f^n(x^n)}{g^n(x^n)} \to 0$$

X に確率密度関数が存在しないとき (Suzuki 2011)

$B_1 := \{\{1\}, \{2, 3, \cdots\}\}$
$B_2 := \{\{1\}, \{2\}, \{3, 4, \cdots\}\}$
$\cdots$
$B_k := \{\{1\}, \{2\}, \cdots, \{k\}, \{k+1, k+2, \cdots\}\}$
$\cdots$

各レベル k で、$x^n = (x_1, \cdots, x_n)$ を $(b_1^{(k)}, \cdots, b_n^{(k)})$ に量子化

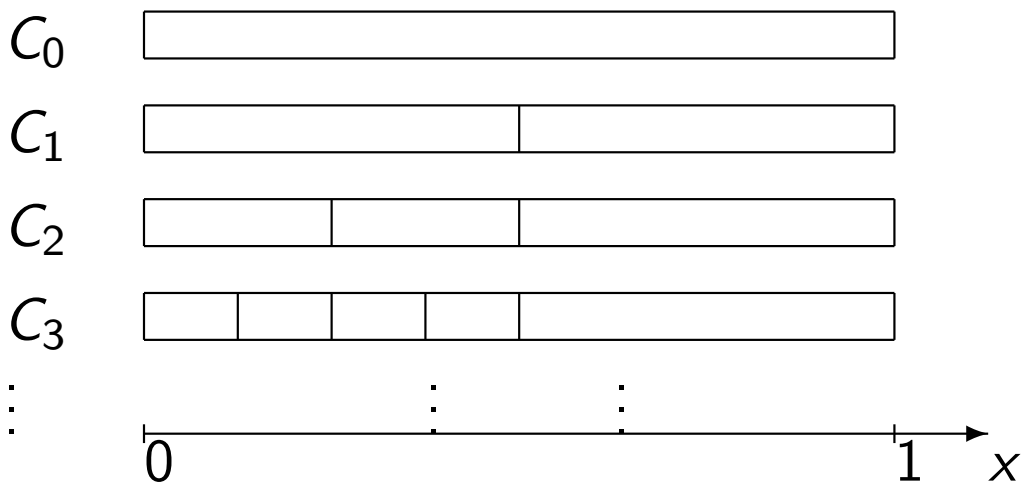
$$\eta(\{k\}) = \frac{1}{k} - \frac{1}{k+1}$$

$$g_k^n(y^n) := \frac{Q_k^n(b_1^{(k)}, \cdots, b_n^{(k)})}{\eta(b_1^{(k)}) \cdots \eta(b_n^{(k)})}$$

$$\sum \omega_k = 1,\ \omega_k > 0,\ g^n(x^n) := \sum_{k=1}^{\infty} \omega_k g_k^n(x^n)$$

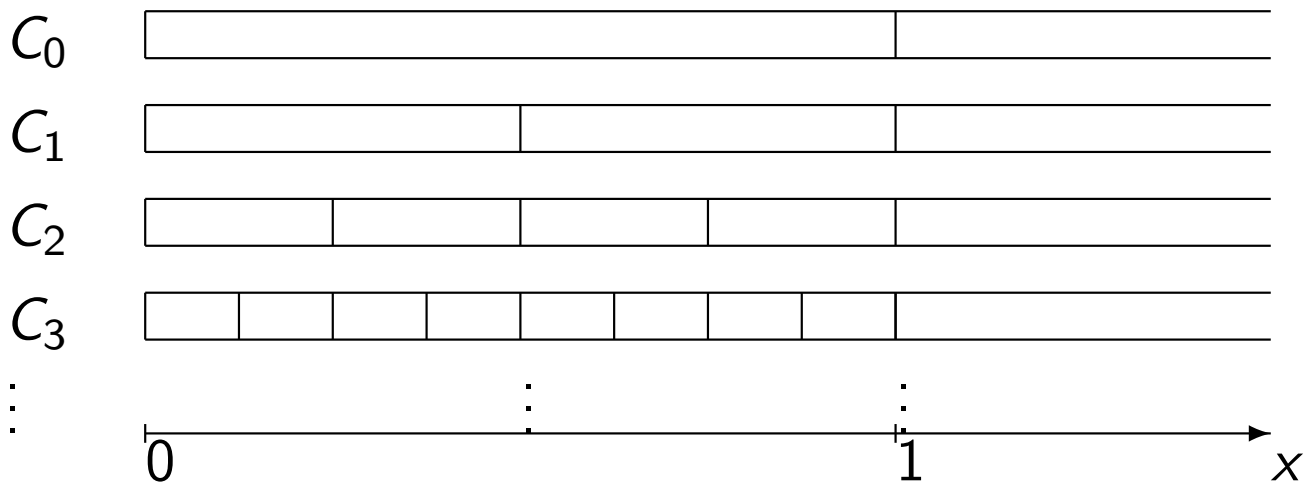
$D(f||f_j) \to 0$ $(j \to \infty)$ にならない例 (その 1)

$$\int_{\frac{1}{2}}^{1} f(x)dx > 0$$



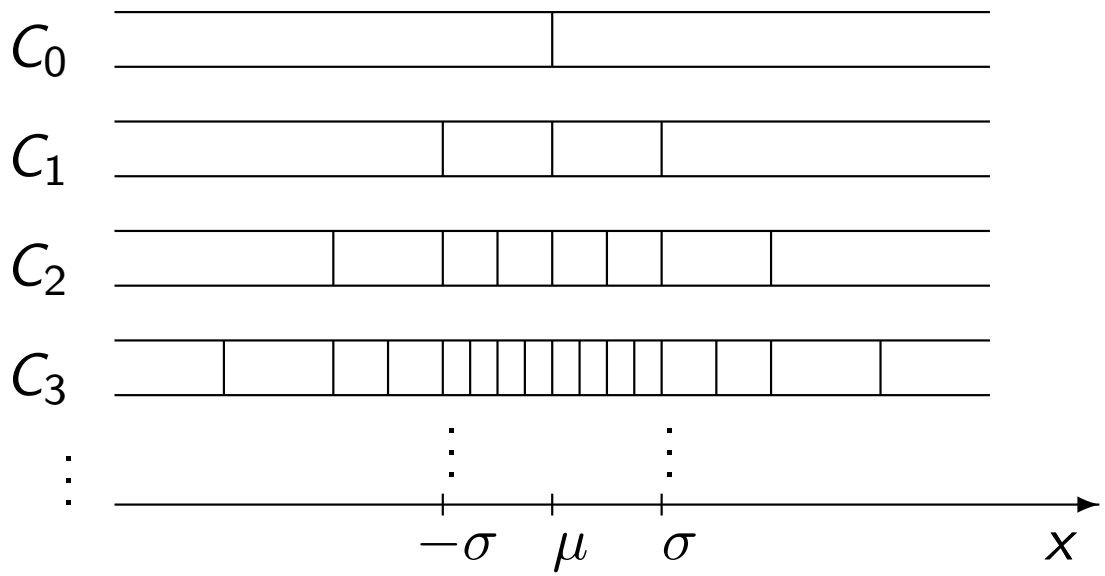
$D(f||f_j) \to 0$ $(j \to \infty)$ にならない例 (その 2)

$$\int_1^\infty f(x)dx > 0$$



ユニバーサル ヒストグラム列 ($D(f||f_j) \to 0$ がつねに満足される)

ユニバーサルなヒストグラム列 $\{C_k\}_{k=0}^{\infty}$



Suzuki 2013

任意の (一般化) 確率密度関数 f について、$n \to \infty$ で確率 1 で、

$$\frac{1}{n} \log \frac{f^n(x^n)}{g^n(x^n)} \to 0$$

2 変数以上の場合

$$g^n_{jk}(x^n, y^n) := \frac{Q^n_{jk}(a^{(j)}_1, \cdots, a^{(j)}_1, b^{(k)}_1, \cdots, b^{(k)}_n)}{\lambda(a^{(j)}_1)\cdots\lambda(a^{(j)}_n)\eta(b^{(k)}_1)\cdots\eta(b^{(k)}_n)}$$

$$\sum_{jk}\omega_{jk} = 1,\ \omega_{jk} > 0,\ g^n(x^n, y^n) := \sum_{k=1}^{\infty}\omega_{jk}g^n_{jk}(x^n, y^n)$$

$g^n(y^n)$, $g^n(z^n)$, $g^n(x^n, y^n)$, $g^n(x^n, z^n)$, $g^n(y^n, z^n)$, $g^n(x^n, y^n, z^n)$ も同様。

$$p_1 g^n(x^n)g^n(y^n)g(z^n), p_2 g^n(x^n)g^n(y^n, z^n), p_3 g^n(y^n)g^n(z^n, x^n),$$

$$p_4 g^n(z^n)g^n(x^n, y^n), p_5\frac{g^n(x^n, y^n)g^n(X, z^n)}{g^n(x^n)}, p_6\frac{g^n(x^n, y^n)g^n(y^n, z^n)}{g^n(y^n)},$$

$$\cdots$$

$$p_{10}\frac{g^n(x^n)g^n(y^n)g^n(x^n, y^n, z^n)}{g^n(x^n, y^n)}, p_{11}g^n(x^n, y^n, z^n)\ ,$$

スコアと構造の評価値の計算

N: ノード数 ($2^N - 1$ 個のスコアを計算)
$M(N)$: 構造の候補数 (例: $M(3) = 11$)

STEP 1: $2^N - 1$ 個の測度のスコアを計算。(11) なら

$$\frac{1}{n}\{-\log p_{11} - \log g^n(x^n) - \log g^n(y^n) - \log g^n(x^n, y^n, z^n) + \log g^n(x^n, y^n)\}$$

	X	Y	(X,Y)	Z	(X,Z)	(Y,Z)	(X,Y,Z)
$-\frac{1}{n}\log g^n(\cdot)$	1.617	1.533	3.249	1.647	3.318	3.290	4.943

STEP 2: $M(N)$ 個の候補のスコアを計算。$N = 3$ なら

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
4.799	4.908	4.852	4.897	4.950	5.006	4.962	4.833	4.890	4.845	4.94

スコア計算のアルゴリズム

入力 $x^n \in A^n$, 出力 $g^n(x^n)$

1. For each $k = 1, \cdots, K$, $g_k^n(x^n) := 0$
2. For each $k = 1, \cdots, K$ and each $a \in A_k$, $c_k(a) := 0$
3. For each $i = 1, \cdots, n$, for each $k = 1, \cdots, K$
   1. Find $a_i \in A_k$ from $x_i \in A$
   2. $g_k^n(x^n) := g_k^n(x^n) - \log \dfrac{c_k(a_i) + 1/2}{i - 1 + |A_k|/2} + \log(\eta x(a_i))$
   3. $c_k(a_i) := c_k(a_i) + 1$
4. $g^n(x^n) := \frac{1}{K} \sum_{k=1}^{K} g_k^n(x^n)$

$$Q_k^n(x^n) = \prod_{i=1}^{n} \frac{c(a_i^{(k)}) + 1/2}{i - 1 + |A|/2}$$

計算量の評価 $\max\{n2^N K, M(N)\} = O(M(N))$

スコア計算

1 測度のスコアで $O(nK)$、$2^N - 1$ 個で $O(n2^N K)$

1 変数の場合:

- サンプル数 n に比例
- $2^N - 1$ 個のスコアを計算
- $a_i^{(1)} \mapsto a_i^{(2)} \mapsto \cdots \mapsto a_i^{(K)}$ が 2 分探索で実現

2 変数以上の場合: $j = k$ として実現すれば、同じ計算量が得られる

$$g^n(x^n, y^n) := \sum_{k=1}^{\infty} \omega_{jk} g_{jk}^n(x^n, y^n)$$

事後確率最大の構造を求める

$O(M(N))$

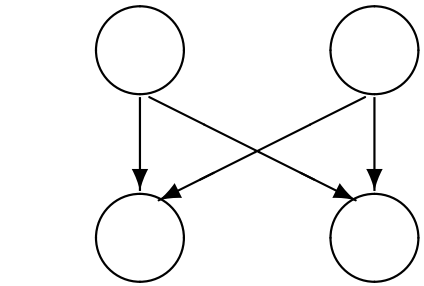
$M(N)$ 個の候補の中での最大値

実験 1

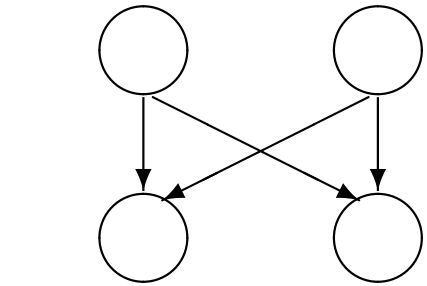
1. $X, Y \in \{0,1\}$ $(X \perp\!\!\!\perp Y)$: 確率 0.5, $U \sim N(x+y,1)$, $V \sim N(x-y,1)$
2. $X, Y \sim N(0,1)$ $(X \perp\!\!\!\perp Y)$, $U, V \in \{0,1\}$ s.t.

$$P(U=1|X+Y=z) = P(V=1|X-Y=z) = \begin{cases} 0, & z < -1 \\ (z+1)/2, & -1 \le z \le 1 \\ 1, & z > 1 \end{cases}$$

(a)



(b)



Bayesian Networks	n	100	200	500	1000	2000
(a)	$g^n(x^n, y^n, u^n, v^n)$	5.009	4.858	4.626	4.616	4.552
4.224	KL divergence	0.785	0.634	0.402	0.392	0.328
	execution time (sec)	1.079	1.276	1.939	4.596	7.047
(b)	$g^n(x^n, y^n, u^n, v^n)$	4.435	4.191	4.002	3.867	3.771
3.372	KL divergence	1.063	0.819	0.630	0.495	0.399
	execution time (sec)	0.601	0.849	1.721	2.582	4.619

実験 2

(1) $X, Y, Z \sim N(0,1)$.

(2)(3)(4) $X, U \sim N(0,1),\ Y = \rho X + \sqrt{1-\rho^2}U,\ Z \sim N(0,1)$.

(5)(6)(7) $X, U, V \sim N(0,1),\ Y = \rho_a X + \sqrt{1-\rho_a^2}U,$
$Y = \rho x + \sqrt{1-\rho_b^2}U$.

(8)(9)(10) $X, U \sim N(0,1),\ Y = \rho_a X + \sqrt{1-\rho_a^2}U,$
$Z = \rho_b X + \sqrt{1-\rho_b^2}V$.

(11) $X, U, V \sim N(0,1),\ Y = \rho_a X + \sqrt{1-\rho_a^2}U,$
$Z = \rho_b X + \rho_c Y + \sqrt{1-\rho_b^2-\rho_c^2}V$.

true structure	differential entropy	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
		score	error	score	error	score	error	score	erro
(1)	4.256816	4.875	0.28	4.645	0.02	4.480	0.00	4.417	0.00
(2)(3)(4)	4.033672	4.699	0.42	4.573	0.12	4.434	0.10	4.350	0.02
(5)(6)(7)	3.810528	4.732	0.34	4.565	0.14	4.385	0.10	4.289	0.02
(8)(9)(10)	3.810528	4.710	0.32	4.498	0.12	4.370	0.06	4.282	0.00
(11)	3.766401	4.731	0.14	4.5431	0.06	4.335	0.02	4.261	0.00

実験 3

R のデータセットについて、実行時間を測定

data.frame	N	data.type	n	time (sec)	time (sec)/2^N
faithful	2	c,d	272	6.08	3.04
quakes	5	c,c,d,d,c	1000	60.77	1.90
attitude	7	d,d,d,d,d,d,d	30	27.66	0.216
longley	7	c,c,c,c,c,c,d	16	44.63	0.349
USJudgeRatings	12	c,c,c,c,c,c,c,c,c,c,c,c	43	1946.63	1.90

N が大きいと、スコアを求める計算が大きいのは、連続でも離散でも同じ

まとめ

離散と連続を区別しない一般的な BN の構造学習の確立

- アルゴリズムの検討 (計算量の評価を含む)
- R による実データでの評価

得られた知見

- 連続でも、ヒストグラムの深さに比例する程度の計算量
- 構造の比較の計算量 $O(M(N))$ が $O(nK2^N)$ より大きい (n 一定の仮定のもとで)

今後の課題

- n, N から、適当な K の値を計算
- K に対して指数的なメモリが必要である問題の解決
- R パッケージの公開
- 実データへのさらなる適用