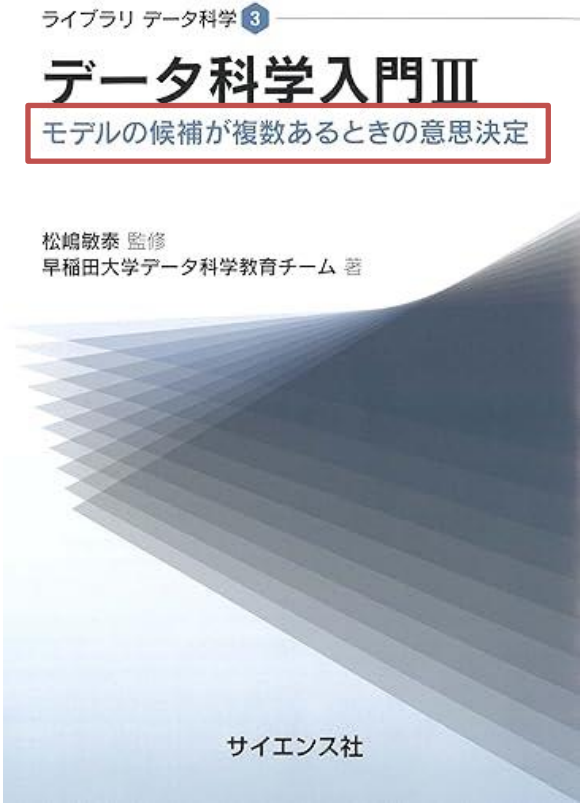
2026年3月11日 IEICE総合大会 企画セッション
AI-6：統一的体系化によるAI・データサイエンス研究の新展望

CENTER FOR DATA SCIENCE

モデルの候補が複数あるときの意思決定

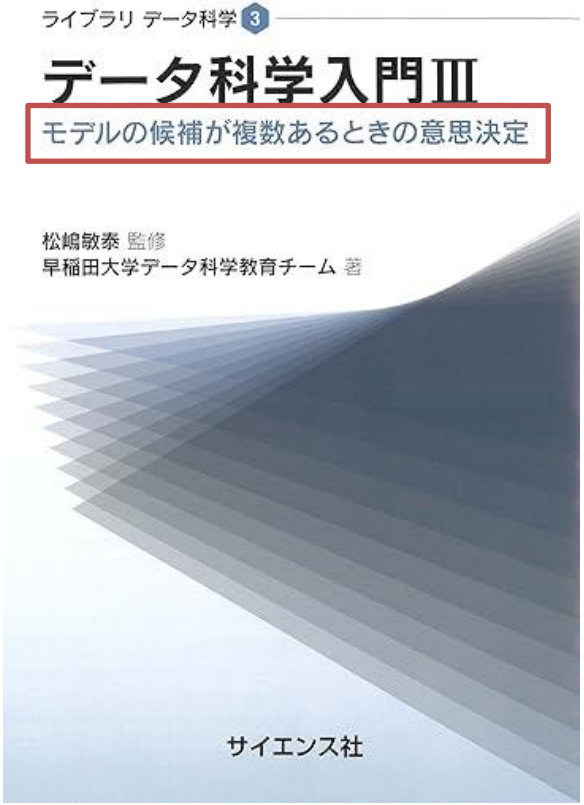
堀井俊佑，松嶋敏泰（早稲田大学）

宣伝



- 講演の前半：教科書の内容
- 講演の後半：教科書に書いていない内容

宣伝



- 講演の前半：教科書の内容
- 講演の後半：教科書に書いていない内容

講演の背景と目的

モデル選択問題

- 複数のモデル候補が存在する状況において、観測データに基づいて適切なモデルを選択する問題

➡ 多くの研究が存在

本講演の内容

- モデル選択の様々な方法を意思決定写像の観点から整理
- モデルを『選択しない』意思決定が最適となる場合について
- これらの枠組みの統計的因果推論の問題への拡張について

モデル選択問題

例1

- 不動産会社が「物件の成約価格」を予測するモデルを作りたい
  - 手元に500件の過去データがある

Y: 価格, X_1: 面積, X_2: 築年数, X_3: 駅距離, …, X_{20}: 近隣学校の評判

モデルの候補

$$m_1: Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$m_2: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$\vdots$$

} モデルの集合$\mathcal{M}$

($\varepsilon \sim N(0, \sigma_\varepsilon^2)$と仮定($\sigma_\varepsilon^2$は未知))

モデル選択問題

例2

- 教育政策の担当者が、小学生の算数の点数に影響している要因を明らかにしたい
  - 手元に1,000件の過去データがある

Y: 算数の点数, X_1: 1日の勉強時間, X_2: 睡眠時間,…, X_{10}: 通っている塾

モデルの候補

$$m_1: Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$m_2: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$\vdots$$

} モデルの集合$\mathcal{M}$

($\varepsilon \sim N(0, \sigma_\varepsilon^2)$と仮定($\sigma_\varepsilon^2$は未知))

モデル選択の意思決定写像としての表現

目的：

設定：　　モデルの候補の集合$\mathcal{M}$

評価基準：

モデル選択の目的

例1

- 不動産会社が「物件の成約価格」を予測するモデルを作りたい
  - 手元に500件の過去データがある

Y: 価格, X_1: 面積, X_2: 築年数, X_3: 駅距離, …, X_{20}: 近隣学校の評判

予測を行うためにモデルを選択したい

モデル選択問題

例2

- 教育政策の担当者が、小学生の算数の点数に影響している要因を明らかにしたい
  - 手元に1,000件の過去データがある

Y: 算数の点数, X_1: 1日の勉強時間, X_2: 睡眠時間,…, X_{10}: 通っている塾

構造を明らかにしたい
（確率的データ生成観測メカニズムの構造を推定したい）

モデル選択の意思決定写像としての表現

目的：　予測を行うためにモデルを選択したい

設定：　モデルの候補の集合$\mathcal{M}$

評価基準：

モデル選択の意思決定写像としての表現

目的：　確率的データ生成観測メカニズムの構造を推定したい

設定：　モデルの候補の集合$\mathcal{M}$

評価基準：

モデル選択問題の設定

モデルの候補

$$m_1: Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$m_2: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$\vdots$$

} モデルの集合$\mathcal{M}$

($\varepsilon \sim N(0, \sigma_\varepsilon^2)$と仮定($\sigma_\varepsilon^2$は未知))

- モデルmとモデルmのもとでのパラメータ$\theta_m = (\boldsymbol{\beta}_m, \sigma_\varepsilon^2)$が定まると、$(\boldsymbol{x}, y)$の分布$p(\boldsymbol{x}, y \mid m, \theta_m)$が定まる
- あるモデル$m^*, \theta_{m^*}^*$が存在して、観測データおよび予測対象のデータはi.i.d.で$p(\boldsymbol{x}, y \mid m^*, \theta_{m^*}^*)$に従う
- 以降、簡単のため$\boldsymbol{x}$とyの組をzとする

モデル選択の意思決定写像としての表現

目的：　予測を行うためにモデルを選択したい

設定：　モデルの候補の集合$\mathcal{M}$

モデルmとパラメータθ_mの元での分布$p(z \mid m, \theta_m)$

$Z_1, \ldots, Z_n, Z_{n+1}$はi.i.d.で$p(z \mid m^*, \theta^*_{m^*})$に従う

評価基準：

モデル選択の意思決定写像としての表現

目的：　確率的データ生成観測メカニズムの構造を推定したい

設定：　モデルの候補の集合$\mathcal{M}$

モデルmとパラメータθ_mの元での分布$p(z \mid m, \theta_m)$

$Z_1, \ldots, Z_n$はi.i.d.で$p(z \mid m^*, \theta^*_{m^*})$に従う

評価基準：

統計的決定理論によるモデル選択の定式化

- 決定関数（意思決定写像）d

例

モデルのみを出力する場合: $d: Z^n \to \mathcal{M}$

モデルとパラメータを出力する場合: $d: Z^n \to (\mathcal{M}, \Theta)$

- 損失関数$\ell(d, Z^n, m, \theta_m)$: 決定関数$d$の「良さ」を評価する関数

モデルのみを出力する場合の0-1損失：

$$\ell(d, Z^n, m, \theta_m) = \begin{cases} 0 & \text{if } d(Z^n) = m \\ 1 & \text{otherwise} \end{cases}$$

例

モデルとパラメータを出力する場合のKL損失：

$$\ell(d, Z^n, m, \theta_m) = \int p(z \mid m, \theta_m) \log \frac{p(z \mid m, \theta_m)}{p(z \mid \hat{m}, \hat{\theta}_{\hat{m}})} \mathrm{d}z$$

（決定関数の出力を$\hat{m}, \hat{\theta}_{\hat{m}}$とおいた）

構造推定を目的としたときの評価基準

- 損失関数として0-1損失を採用

$$\ell(d, Z^n, m, \theta_m) = \begin{cases} 0 & \text{if } d(Z^n) = m \\ 1 & \text{otherwise} \end{cases}$$

- 危険関数: 損失関数をデータに関して期待値をとったもの

$$\begin{aligned} R(d, m, \theta_m) &= \mathbb{E}[\ell(d, Z^n, m, \theta_m)] \\ &= \int \ell(d, z^n, m, \theta_m) p(z^n \mid m, \theta_m) \mathrm{d}z^n \\ &= \Pr(d(Z^n) \neq m) \end{aligned}$$

- 一般的に、任意のm, θ_mに対して危険関数を最小化する決定関数は存在しない

⇒ 事前分布$p(m), p(\theta_m \mid m)$を導入

構造推定を目的としたときの評価基準

- ベイズ危険関数：危険関数を事前分布で期待値をとったもの

$$BR(d) = \mathbb{E}[R(d, m, \theta_m)]$$
$$= \sum_{m \in \mathcal{M}} p(m) \int R(d, m, \theta_m) p(\theta_m \mid m) \mathrm{d}m$$

- ベイズ危険関数は決定関数dのみの関数

⇒ dに関する最適化を考えることが可能

$$d^* = \arg\min_d BR(d)$$

- 0-1損失を考えた場合、事後確率最大のモデルを出力するのが最適

$$d^*(Z^n) = \arg\max_{m \in \mathcal{M}} p(m \mid Z^n)$$

モデル選択の意思決定写像としての表現

目的：	確率的データ生成観測メカニズムの構造を推定したい
設定：	モデルの候補の集合$\mathcal{M}$
	モデルmとパラメータθ_mの元での分布$p(z \mid m, \theta_m)$
	$Z_1, \ldots, Z_n$はm, θ_mのもとi.i.d.で$p(z \mid m, \theta_m)$に従う
	mとθ_mの事前分布$p(m), p(\theta_m \mid m)$
評価基準：	0-1損失に対するベイズ危険関数最小化

最適な意思決定写像：事後確率を最大にするモデルを出力

モデルの事後確率とBIC

- モデルの事前分布として、一様分布を仮定

$$p(m) = \frac{1}{|\mathcal{M}|}$$

- 事後確率の最大化は周辺尤度の最大化に帰着

$$\arg\max_{m \in \mathcal{M}} p(m \mid Z^n) = \arg\max_{m \in \mathcal{M}} p(Z^n \mid m)$$

- 周辺尤度の計算式

$$p(z^n \mid m) = \int p(z^n \mid m, \theta_m) p(\theta_m \mid m) \mathrm{d}\theta_m$$

積分が困難な場合が多い

- 最尤推定量の漸近正規性など、いくつかの条件のもとで

$$\frac{1}{n} \log p(z^n \mid m) \approx \frac{1}{n} \log p\left(z^n \mid m, \hat{\theta}_{m,\mathrm{ML}}\right) - \frac{d_m}{2n} \log n$$

➡ BIC基準

予測を目的としたときの評価基準

- Z_{n+1}は確率変数 ⇒ 確率的な評価が必要
  - $\mathbb{E}[Z_{n+1}]$を推定対象とする
  - $p(z \mid m^*, \theta^*_{m^*})$を推定対象とする

- 推定モデルを$\hat{m}$、推定パラメータを$\hat{\theta}_{\hat{m}}$とする（決定関数dの出力）

⇒ 推定分布：$p(z \mid \hat{m}, \hat{\theta}_{\hat{m}})$

- 損失関数としてKLダイバージェンスを採用

$$\ell(d, Z^n, m, \theta_m) = \int p(z \mid m, \theta_m) \log \frac{p(z \mid m, \theta_m)}{p(z \mid \hat{m}, \hat{\theta}_{\hat{m}})} \mathrm{d}z$$

予測を目的としたときの評価基準

- KLダイバージェンスを展開

$$\ell(d, Z^n, m, \theta_m) = \int p(z \mid m, \theta_m) \log \frac{p(z \mid m, \theta_m)}{p(z \mid \hat{m}, \hat{\theta}_{\hat{m}})} \mathrm{d}z$$

$$= \underline{\int p(z \mid m, \theta_m) \log p(z \mid m, \theta_m) \mathrm{d}z} - \int p(z \mid m, \theta_m) \log p(z \mid \hat{m}, \hat{\theta}_{\hat{m}}) \mathrm{d}z$$

$\hat{m}, \hat{\theta}_{\hat{m}}$に依存しない

- 損失関数の書き換え

$$\ell'(d, Z^n, m, \theta_m) = -\int p(z \mid m, \theta_m) \log p(z \mid \hat{m}, \hat{\theta}_{\hat{m}}) \mathrm{d}z$$

予測を目的としたときの評価基準

- 危険関数：損失関数をデータに関して期待値をとったもの

$$R(d, m, \theta_m) = \mathbb{E}[\ell'(d, Z^n, m, \theta_m)]$$

$$= \int \ell'(d, z^n, m, \theta_m) p(\, z^n \mid m, \theta_m \,) \mathrm{d}z^n$$

- 一般的に、任意のm, θ_mに対して危険関数を最小化する決定関数は存在しない
- 一つの方法：事前分布を設定してベイズ危険関数最小化
- ここでは別の方法を紹介
- 真のモデル$m^*, \theta^*_{m^*}$が存在すると仮定し、以下の最小化を考える

$$\int \ell'(d, z^n, m^*, \theta^*_{m^*}) p(\, z^n \mid m^*, \theta^*_{m^*} \,) \mathrm{d}z^n$$

予測を目的としたときの評価基準

- 最小化したい目的関数：

$$\int \ell'(d, z^n, m^*, \theta^*_{m^*}) p(\, z^n \mid m^*, \theta^*_{m^*}\,) \mathrm{d}z^n$$

- 2つの困難性：
  - 真の分布$p(\, z \mid m^*, \theta^*_{m^*}\,)$は未知

  ⇒ 推定量で近似
  - mとθ_mの同時最適化が困難

  ⇒ θ_mについては最尤推定量$\hat{\theta}_m^{(\mathrm{ML})}$を使うことにする

予測を目的としたときの評価基準

- 修正した目的関数：

$$\int \ell'\left(m, \hat{\theta}_{m,\mathrm{ML}}, m^*, \theta^*_{m^*}\right) p\left(z^n \mid m^*, \theta^*_{m^*}\right) \mathrm{d}z^n$$

- 漸近不偏推定量：

$$-\frac{1}{n}\sum_{i=1}^{n} \log p\left(Z_i \mid \hat{m}, \hat{\theta}^{(\mathrm{ML})}_{\hat{m}}\right) + \frac{d_m}{n}$$

AIC基準

モデル選択の意思決定写像としての表現

目的：	予測を行うためにモデルを選択したい
設定：	モデルの候補の集合$\mathcal{M}$
	モデルmとパラメータθ_mの元での分布$p(z \mid m, \theta_m)$
	$Z_1, \ldots, Z_n, Z_{n+1}$はi.i.d.で$p(z \mid m^*, \theta^*_{m^*})$に従う
評価基準：	KL損失に対する危険関数の漸近不偏推定量最小化

最適な意思決定写像：AIC基準を最小にするモデルを出力

間接予測と直接予測

- 目的が予測である場合、必ずしもモデル選択は必須ではない [松嶋, 98]

- モデルを選択してから予測 ⇒ **間接予測**
- モデルを選択せずに予測 ⇒ **直接予測**

モデル選択の意思決定写像としての表現

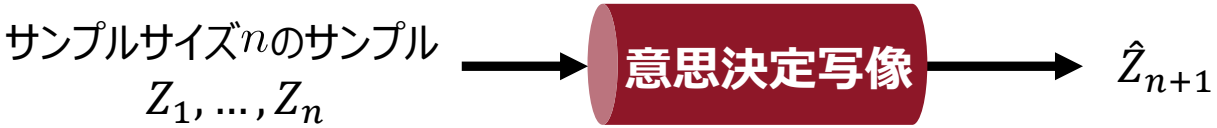
目的：　予測を行いたい

設定：　モデルの候補の集合$\mathcal{M}$

モデルmとパラメータθ_mの元での分布$p(z \mid m, \theta_m)$

$Z_1, \ldots, Z_n, Z_{n+1}$は$m, \theta_m$のもとi.i.d.で$p(z \mid m, \theta_m)$に従う

評価基準：



統計的決定理論による予測問題の定式化

- 決定関数（意思決定写像）$d : Z^n \to Z_{n+1}$

- 損失関数$\ell(d, Z^n, m, \theta_m)$: 二乗誤差損失

$$\ell(d, Z^n, m, \theta_m) = \mathbb{E}[(d(Z^n) - Z_{n+1})^2]$$

$$= \int (d(Z^n) - z_{n+1})^2 p(z_{n+1} \mid m, \theta_m) \mathrm{d}z_{n+1}$$

- 危険関数：損失関数をデータに関して期待値をとったもの

$$R(d, m, \theta_m) = \mathbb{E}[\ell(d, Z^n, m, \theta_m)]$$

$$= \int \ell(d, z^n, m, \theta_m) p(z^n \mid m, \theta_m) \mathrm{d}z^n$$

統計的決定理論による予測問題の定式化

- ベイズ危険関数：危険関数を事前分布で期待値をとったもの

$$BR(d) = \mathbb{E}[R(d, m, \theta_m)]$$
$$= \sum_{m \in \mathcal{M}} p(m) \int R(d, m, \theta_m) p(\theta_m \mid m) \mathrm{d}m$$

- 二乗誤差損失を仮定した場合の最適なd

$$d^*(z^n) = \int z_{n+1} \cdot p(z_{n+1} \mid z^n) \mathrm{d}z_{n+1}$$

$$p(z_{n+1} \mid z^n) = \sum_{m \in \mathcal{M}} p(m \mid z^n) \int p(z_{n+1} \mid m, \theta_m) p(\theta_m \mid m, z^n) \mathrm{d}\theta_m$$

（事後）予測分布

直接予測の意思決定写像としての表現

目的：　予測を行いたい

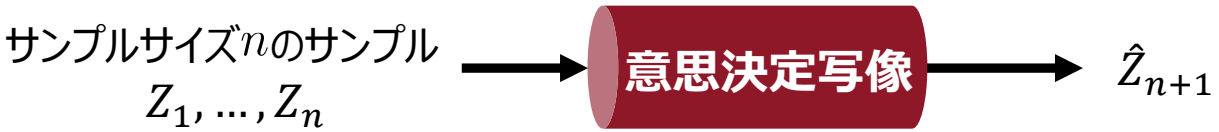
設定：　モデルの候補の集合$\mathcal{M}$

モデルmとパラメータθ_mの元での分布$p(z \mid m, \theta_m)$

$Z_1, \ldots, Z_n, Z_{n+1}$は$m, \theta_m$のもとi.i.d.で$p(z \mid m, \theta_m)$に従う

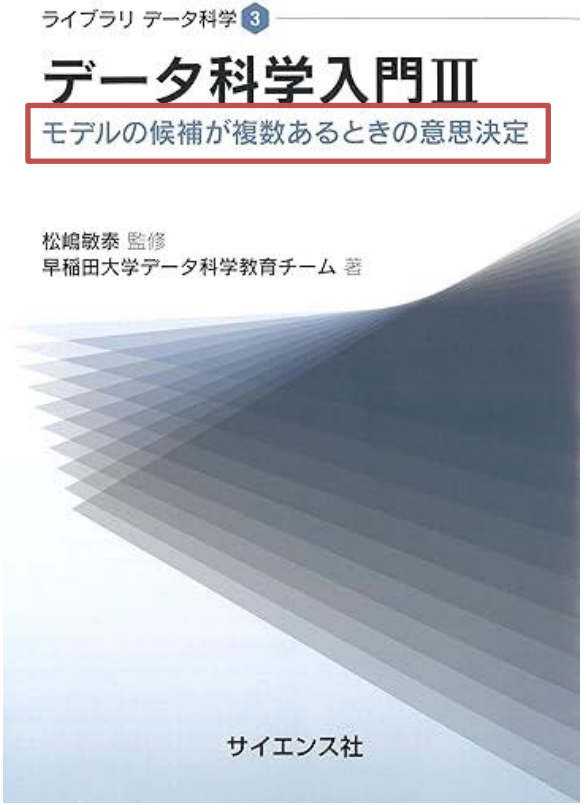
mとθ_mの事前分布$p(m), p(\theta_m \mid m)$

評価基準：　二乗誤差損失に対するベイズ危険関数最小化



最適な意思決定写像：事後予測分布の期待値を出力

宣伝



- 講演の前半：教科書の内容
- 講演の後半：教科書に書いていない内容

構造的因果モデルの概要

- 構造的因果モデル：因果関係を数学的に記述するフレームワーク

- 2つの構成要素：
  1. 構造方程式モデル：変数間の因果関係を方程式で記述
  2. 因果ダイアグラム：因果構造を有向グラフで視覚化

- 特徴：
  - 因果関係の方向とメカニズムを明示的に表現
  - 「介入」を数学的に定義可能

構造方程式モデルの考え方

- 構造方程式：各変数がどの変数から影響を受けるかを、方程式で明示的に記述

- 例：学生の試験成績に関する構造方程式
  - X：補習を受けた時間
  - T：宿題の量
  - Y：試験の点数

- 変数間の局所的な因果関係
  - 補習時間が長いほど、宿題も多くこなす（$X \to T$）
  - 補習時間が長いほど、試験の点数も高い（$X \to Y$）
  - 宿題が多いほど、試験の点数も高い（$T \to Y$）

構造方程式モデルの考え方

- 構造方程式

$$X = U_X$$
$$T = 0.5X + U_T$$
$$Y = 0.7X + 0.4T + U_Y$$

※ すべての変数は標準化後の値を表すとする

錯乱項（U_X, U_T, U_Y）：

- 観測されない外生的な要因
- 互いに独立に分布

補足：通常、関数型やパラメータは観測データから推定する

構造方程式モデルの特徴

- 錯乱項の値が決まると、すべての変数の値が確定的に決まる

例：$U_X = 0.5, U_T = 0.75, U_Y = 0.75$の場合

$$X = 0.75$$
$$T = 0.5 \times 0.5 + 0.75 = 1.0$$
$$Y = 0.7 \times 0.5 + 0.4 \times 1.0 + 0.75 = 1.5$$

- 観測される変数は、その背後にある錯乱項によって決定される

構造方程式モデルと周辺分布・条件付き分布

- 構造方程式は各変数の周辺分布・条件付き分布を規定する

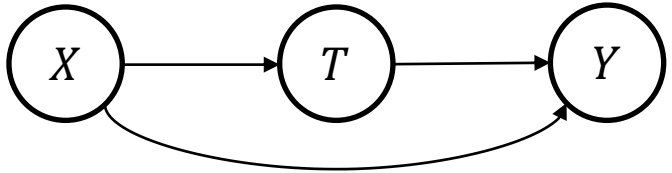
$$X = U_X \quad \Rightarrow \quad p(x)\text{を規定}$$
$$T = 0.5X + U_T \quad \Rightarrow \quad p(t \mid x)\text{を規定}$$
$$Y = 0.7X + 0.4T + U_Y \quad \Rightarrow \quad p(y \mid x, t)\text{を規定}$$

- 右辺の変数を条件としたときの、左辺の変数の条件付き分布が定まる

因果ダイアグラム

- 因果ダイアグラム：構造方程式を有向グラフで表現
  - ノード：変数
  - 矢線：因果関係の方向（右辺の変数から左辺の変数へ）

$$X = U_X$$
$$T = 0.5X + U_T$$
$$Y = 0.7X + 0.4T + U_Y$$



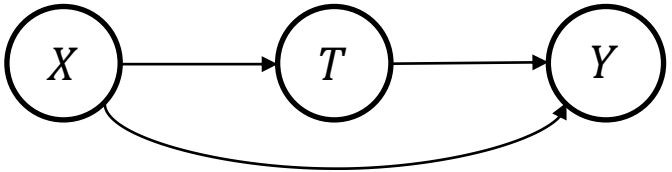
- 因果ダイアグラムは構造方程式から一意に定まる
- 因果ダイアグラムは非巡回であることを仮定
  - 因果ダイアグラムはDAG（Directed Acyclic Graph）

同時分布の因子分解構造

- 因果ダイアグラムの構造が同時分布の因子分解構造を与える

$$p(x_1, \ldots, x_p) = \prod_{j=1}^{p} p(x_j \mid \mathrm{Pa}(x_j))$$

$\mathrm{Pa}(x_j)$: x_jの親ノード変数の集合



$$p(x, t, y) = p(x)p(t \mid x)p(y \mid x, t)$$

自律性の仮定

- 構造方程式の一部を変化させたとしても、その他の構造方程式に影響はない

元の構造方程式：

$$X = U_X$$
$$T = 0.5X + U_T$$
$$Y = 0.7X + 0.4T + U_Y$$

Tの構造方程式を$T = 2$に変更

$$X = U_X$$
$$T = 2$$
$$Y = 0.7X + 0.4T + U_Y$$

- この仮定により、「構造方程式の一部を変更した世界」の話ができる

介入

- 介入：系の外部から変数の値を強制的に設定すること

- 記法：$\mathrm{do}(T = t)$
  - 「Tの値をtに設定する」という介入を表す

- 「もしTの値がtだったら」何が起こるかに興味がある
  - 宿題の量を全員0にしたら？$\Rightarrow \mathrm{do}(T = 0)$
  - 補習の量を全員0にしたら？$\Rightarrow \mathrm{do}(X = 0)$

介入分布

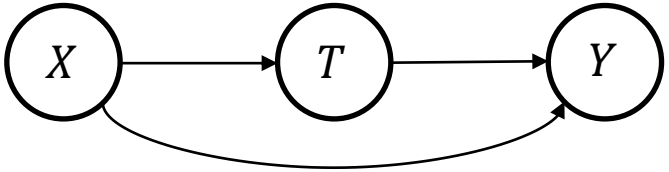
- 介入分布：$\mathrm{do}(T = t)$という介入を行ったときの分布。$p_{\mathrm{do}(T=t)}(\cdot)$により表す。

- 変数$T, Y, X_1, \ldots, X_p$に対し、 $\mathrm{do}(T = t)$の介入分布は：

$$p_{\mathrm{do}(T=t)}(y, x_1, \ldots, x_p) = p(y \mid \mathrm{Pa}(y)) \prod_{j=1}^{p} p(x_j \mid \mathrm{Pa}(x_j))$$

T以外の変数について、親ノードを条件とした条件付き分布の積で表される

介入分布



構造方程式：

$$X = U_X$$
$$T = 0.5X + U_T$$
$$Y = 0.7X + 0.4T + U_Y$$

介入分布：

$$p_{\mathrm{do}(T=t)}(x, y) = p(y \mid x, t)p(x)$$

介入分布の周辺分布：

$$p_{\mathrm{do}(T=t)}(y) = \int p(y \mid x, t)p(x)\mathrm{d}x$$

介入分布の期待値：

$$\mathbb{E}_{\mathrm{do}(T=t)}[Y] = \int y \cdot p_{\mathrm{do}(T=t)}(y)\,\mathrm{d}y$$

$T = t$としたときのYの期待値

平均因果効果

- 2値変数TのYに対する平均因果効果

$$\mathrm{ACE} = \mathbb{E}_{\mathrm{do}(T=1)}[Y] - \mathbb{E}_{\mathrm{do}(T=0)}[Y]$$

- $T = 1$に介入した場合のYの期待値と$T = 0$に介入した場合のYの期待値の差
- 条件付き期待値の差($\mathbb{E}[\,Y \mid T = 1\,] - \mathbb{E}[\,Y \mid T = 0\,]$)との違いに注意

介入時の分布や期待値の推定

- $p_{\mathrm{do}(T=t)}(y), \mathbb{E}_{\mathrm{do}(T=t)}[Y]$は、因果ダイアグラムや構造方程式が既知ならば計算可能

- 因果ダイアグラムや構造方程式が未知の場合は、データから推定する必要がある
- 問題設定は様々：
  - 因果ダイアグラム：既知、構造方程式の関数型：既知、構造方程式のパラメータ：未知
  - 因果ダイアグラム：既知、構造方程式の関数型：未知、構造方程式のパラメータ：未知
  - 因果ダイアグラム：未知、構造方程式の関数型：未知、構造方程式のパラメータ：未知

  ⇒ 意思決定写像の設定

因果探索

- 因果ダイアグラムが未知のときに、因果ダイアグラムを推定する問題は**因果探索**とよばれる
- 代表例：LiNGAM (Linear Non-Gaussian Acyclic Model) [Shimizu et al., 2006]
  - 構造方程式は線形を仮定

$$X_i = \sum_{j<i} b_{ij} X_j + \varepsilon_i$$

  - ε_iは非ガウス分布を仮定
  - このとき、因果構造が正しければ残差ε_iが他の変数と独立

➡ 独立性を最大にする構造を探索

LiNGAMの意思決定写像としての表現

目的：	因果ダイアグラムを推定したい
設定：	グラフ構造はDAGである
	未観測の変数は存在しない
	構造方程式は線形で、パラメータは未知
	錯乱項の分布は非正規分布である
評価基準：	残差と説明変数の独立性最大化

因果探索の目的

- 因果推論をする上で因果探索をする目的は、介入分布やその期待値を推定すること
- 介入分布の推定精度を考慮した因果探索は？

[D. Janzin, B. Schölkopf, 2013]

介入分布推定を目的とした因果探索

- 因果ダイアグラムGとパラメータθ_Gが定まると、介入分布が定まる

$$p_{\mathrm{do}(T=t)}(y \mid G, \theta_G)$$

- 真の介入分布に対するKLダイバージェンス：

$$KL(p^*_{\mathrm{do}(T=t)}(y) || p_{\mathrm{do}(T=t)}(y \mid G, \theta_G))$$

- 最適な因果ダイアグラムとパラメータ：

$$\hat{G}, \hat{\theta}_G = \arg\min_{G, \theta_G} KL(p^*_{\mathrm{do}(T=t)}(y) || p_{\mathrm{do}(T=t)}(y \mid G, \theta_G))$$

具体的なアルゴリズムは様々なものが提案されている

介入分布推定を目的とした因果探索の意思決定写像としての表現

目的：	介入分布を推定したい
設定：	グラフ構造はDAGである
	未観測の変数は存在しない
	構造方程式は関数形は既知で、パラメータは未知
	錯乱項の分布は分布形は既知で、パラメータは未知
評価基準：	真の介入分布とのKLダイバージェンス最小化

因果ダイアグラムの推定は必要？

- 予測問題：予測が目的ならモデルを1つに固定する必要はない

- 因果推論：介入分布やACEの推定が目的ならモデルを1つに固定する必要は無いのでは？

モデルを1つに固定しない介入分布・ACE推定（直接介入分布推定）
（[H., Suko, 2019], [H., 2021]）

統計的決定理論による介入分布推定の定式化

- 決定関数（意思決定写像）$d : D^n \to \hat{p}_{\mathrm{do}(T=t)}(y)$
- 損失関数：$\ell(d, D^n, G, \theta_G) = KL(p_{\mathrm{do}(T=t)}(y \mid G, \theta_G) || \hat{p}_{\mathrm{do}(T=t)}(y))$
- 危険関数：$R(d, G, \theta_G) = \mathbb{E}_{D^n}[\ell(d, D^n, G, \theta_G)]$
- ベイズ危険関数：$BR(d) = \mathbb{E}_{G,\theta_G}[R(d, G, \theta_G)]$

ベイズ危険関数を最小化する決定関数が最適：

$$d^*(D^n) = p_{\mathrm{do}(T=t)}(y \mid D^n)$$

$$p_{\mathrm{do}(T=t)}(y \mid D^n) = \sum_G p(G \mid D^n) \int p_{\mathrm{do}(T=t)}(y \mid G, \theta_G) p(\theta_G \mid G, D^n) \mathrm{d}\theta_G$$

名付けるなら介入事後予測分布？

直接介入分布の意思決定写像としての表現

目的：　介入分布を推定したい

設定：　グラフ構造はDAGである

未観測の変数は存在しない

構造方程式は関数形は既知で、パラメータは未知

錯乱項の分布は分布形は既知で、パラメータは未知

G, θ_Gの事前分布$p(G), p(\theta_G \mid G)$

評価基準：　介入分布とのKLダイバージェンスを損失関数としたときの

ベイズ危険関数の最小化

最適な意思決定写像：介入事後予測分布を出力

直接介入分布の意思決定写像としての表現

目的：　介入分布を推定したい

設定：　グラフ構造はDAGである

未観測の変数は存在しない

構造方程式は関数形は既知で、パラメータは未知

錯乱項の分布は分布形は既知で、パラメータは未知

G, θ_Gの事前分布$p(G)\ p(\theta_G \mid G)$

評価基準：　介入分布とのKL

ベイズ危険関数

積分計算が難しかったり、
モデルの重み付けが難しかったりするので、
色々な工夫が必要

サンプルサイズnのサンプル → 意思決定 → 推定介入分布

最適な意思決定写像：介入事後予測分布を出力

まとめ

- モデルが未知な設定における様々な意思決定の問題を意思決定写像の視点から整理
- 意思決定写像のフレームワークは研究においても有用
- 細かい設定や、最適な意思決定 or 近似的に最適な意思決定を構築するアルゴリズムなど、考慮すべき点は様々

参考文献

- 松嶋敏泰 (監修), 早稲田大学データ科学教育チーム (著), 「データ科学入門Ⅲ：モデルの候補が複数あるときの意思決定」（サイエンス社）
- 松嶋敏泰, “帰納・演繹推論と予測－決定理論による学習モデル－”, 情報論的学習理論ワークショップ（IBIS）98
- 黒木学, 「構造的因果モデルの基礎」, (共立出版)
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. “A Linear Non-Gaussian Acyclic Model for Causal Discovery.” Journal of Machine Learning Research (JMLR), 7: 2003–2030.
- JANZING, D., BALDUZZI, D., GROSSE-WENTRUP, M. O. R. I. T. Z., & SCHÖLKOPF, B. ”Quantifying causal influences,” The Annals of Statistics, 41(5), 2324-2358.
- Horii, S., & Suko, T. “A note on the estimation method of intervention effects based on statistical decision theory,” In 2019 53rd Annual Conference on Information Sciences and Systems (CISS) (pp. 1-6)
- Horii, S. “Bayesian model averaging for causality estimation and its approximation based on gaussian scale mixture distributions,“ In International Conference on Artificial Intelligence and Statistics (pp. 955-963).