1. IPWの拡張

● 多種類介入へのIPWの適用

■ 提案する損失関数の式

$$L_{IPW} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{p(\boldsymbol{z} = \boldsymbol{z}_i | \boldsymbol{x} = \boldsymbol{x}_i)}$$

一般化傾向スコアを
多種類介入に拡張

組み合わせ数が2^K個と、
指数的に増加するため

介入種類が多いと、一般化傾向スコアを求めるのは難しい

↓

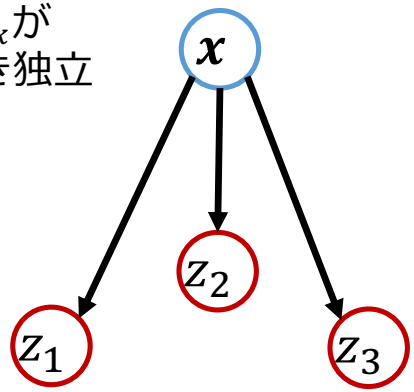
最終的に「結果y」の予測精度が高くなるような、
一般化傾向スコアの評価方法を検討する

2.一般化傾向スコアの評価

- 因果探索による評価

（介入が3種類（$\boldsymbol{z} = \{z_1, z_2, z_3\}$）の場合）

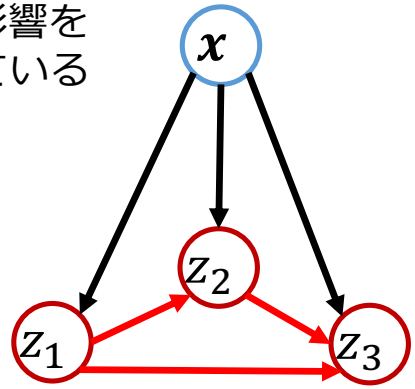
各介入z_kが
条件付き独立



$p(\boldsymbol{z}|\boldsymbol{x}) = p(z_1|\boldsymbol{x}) \cdot p(z_2|\boldsymbol{x}) \cdot p(z_3|\boldsymbol{x})$

因果探索

各介入z_kが影響を
及ぼし合っている

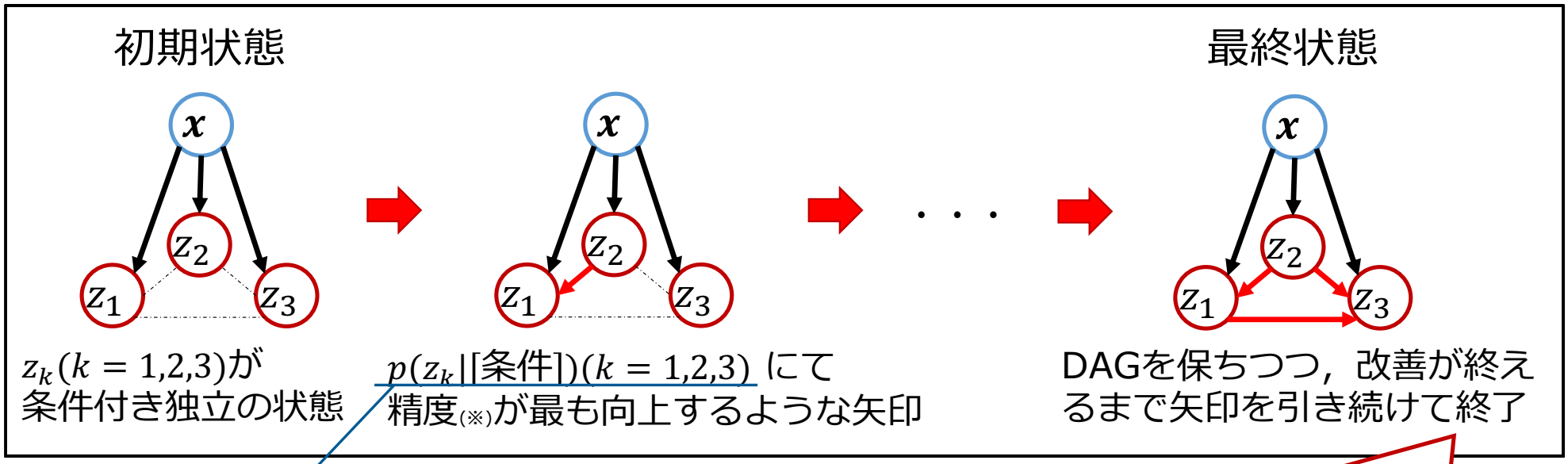


$p(\boldsymbol{z}|\boldsymbol{x}) = p(z_1|\boldsymbol{x}) \cdot p(z_2|\boldsymbol{x}, z_1) \cdot p(z_3|\boldsymbol{x}, z_1, z_2)$

- 因果関係を明確にできれば、**より厳密な**一般化傾向スコアの評価になる
- グラフの種類は非常に多く、全探索は難しいため、**貪欲アルゴリズム**で探索

2. 一般化傾向スコアの評価

- 因果探索の貪欲アルゴリズム
  （介入が3種類（$\boldsymbol{z} = \{z_1, z_2, z_3\}$）の場合）



研究２ 目次

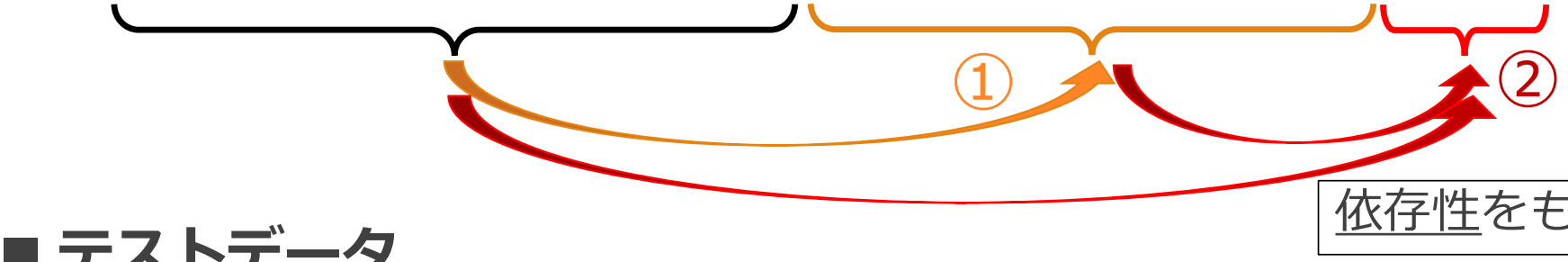
- はじめに
- 関連研究
- 提案手法
- 数値実験

人工データの生成方法

- 全体のコンセプト
  - **学習データ & 検証用データ**

ID	x_1	x_2	…	x_5	z_1	z_2	…	z_5	y
1	-0.34	2.10	…	1.21	1	0	…	1	31.1
2	1.12	-0.05	…	3.34	1	1	…	0	25.2
3	1.67	-0.11	…	2.21	0	0	…	1	11.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

※表はイメージ



  - **テストデータ**
    - 「介入z」を、①に依存せず一様的に生成する

人工データの生成方法

- 事前準備
  - ■ **重みベクトル** (w_x, w_z) を設定
    - 線形結合により、**ベクトル** (x, z) を**スカラー** (x_s, z_s) に変換する
    - 「健康の度合い」「介入の度合い」を表しやすくする

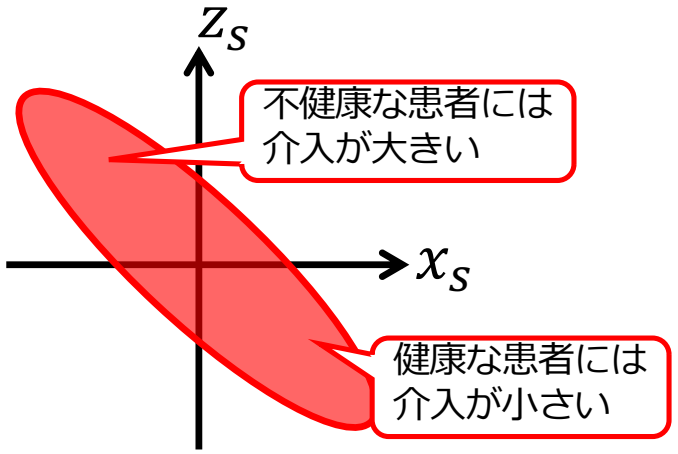
$$x_s = w_x{}^T x, \quad z_s = w_z{}^T z$$

健康の度合い
（大きいほど健康）

介入の度合い
（大きいほど強い）

「重み」ベクトル

目指す人工データの分布



人工データの生成方法

- 確率的生成の方法
  - スカラー (x_s, z_s) を活用して、以下の手順で各変数を確率的に生成する
  - $\{[\boldsymbol{x}, \boldsymbol{z}], y\}$ を、実験用データとして使用する

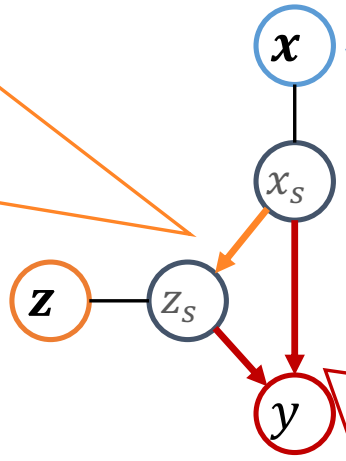
1.正規乱数による生成

2.状態(x_s)に対し、介入 $\boldsymbol{z}(z_s)$を生成

$$p(\boldsymbol{z}|\boldsymbol{x}) \propto f(z_s - \frac{\boldsymbol{w}_z{}^T \mathbf{1}}{1 + \exp(\alpha \cdot x_s)})$$

・前ページのような分布になる
・$f(\cdot)$：平均0, 分散σ^2のガウス密度関数
・$\alpha\ (> 0)$: 介入戦略の極端さ
・σ^2：意思決定のブレの大きさ

↑ テストデータでは、完全ランダムに生成



3.状態(x_s),介入(z_s)に対し、結果 y を生成

$$y = 10 \cdot \frac{1}{1+\exp(\beta_1 \cdot x_s)} \cdot \frac{1}{1+\exp(\beta_2 \cdot z_s)} + \varepsilon$$

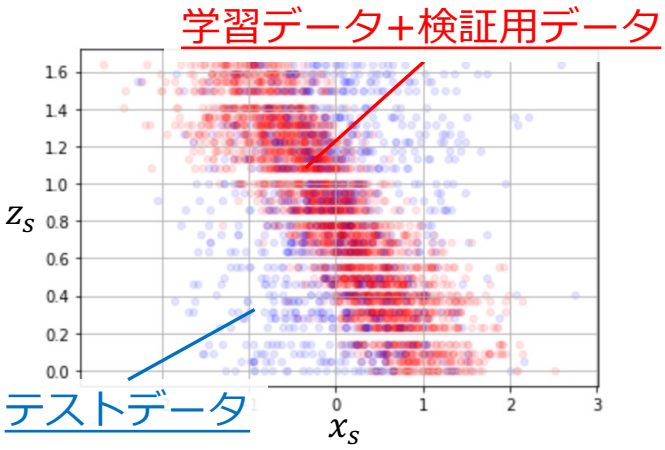
・不健康患者に介入なし → yが大きい
・$\beta_1\ (> 0)$: 「状態」の結果への敏感さ
・$\beta_2\ (> 0)$: 「介入」の結果への敏感さ
・ε : 正規乱数

実験の設定

Tokyo Tech

●タスク

- ■ 『$\boldsymbol{x} \in R^5$と $\boldsymbol{z} \in \{0,1\}^5$を特徴量として、 $y \in R$ を高精度に予測する』
  - バイアスのある「**学習データ**」「**検証用データ**」 で学習
  - バイアスのない「**テストデータ**」で予測精度を評価
  - 学習：検証用：テスト = 1,000：1,000：1,000 (件)
  - 100セットの人工データそれぞれで評価し、平均をとる



●モデルの前提

- ■ 各介入z_k の確率出力モデル：ロジスティック回帰
- ■ y を予測するモデル：Random Forest

実験の設定

- 評価指標
  - **決定係数** (R^2) ： 1に近いほど高精度
- 比較手法（ 一般傾向スコア$p(\boldsymbol{z}|\boldsymbol{x})$ の評価方法）
  1. **「Naive」** ： 考慮しない方法（$p(\boldsymbol{z}|\boldsymbol{x})=1$）
  2. **「多クラス分類」** ： $32(=2^5)$クラスのロジスティック回帰
  3. **「条件付き独立」** ： $p(\boldsymbol{z}|\boldsymbol{x}) = \prod_{k=1}^{5} p(z_k|\boldsymbol{x})$
  4. **「提案手法」** ： 因果探索による評価

実験結果

- 予測精度
  - あるパラメータの人工データにおける、検証用データとテストデータの評価
    - パラメータ：$\alpha = 2, \sigma^2 = 0.04, \beta_1 = 4, \beta_2 = 1$

バイアスあり　バイアスなし

	検証用データ(R^2)	テストデータ(R^2)
Naive	0.8902 ± 0.0530	0.7508 ± 0.1949
多クラス分類	0.8966 ± 0.0464	0.7933 ± 0.1438
条件付き独立	0.8951 ± 0.0485	0.7944 ± 0.1380
提案手法	0.8970 ± 0.0464	0.8004 ± 0.1336

※対応のあるt検定で1%有意

- テストデータについて、**提案手法が最高精度**
- 検証用データにおける精度もほぼ不変

実験結果

- 予測精度
  - 様々なパラメータの人工データにおける、テストデータの評価

バイアスなし

  - ベースのパラメータ：$\alpha = 2, \sigma^2 = 0.04, \beta_1 = 4, \beta_2 = 1$

	$\alpha = 20$	$\alpha = 0.5$	$\sigma^2 = 0.16$	$\sigma^2 = 0.01$	$\beta_1 = 7, \beta_2 = 2$	$\beta_1 = 0.4, \beta_2 = 0.1$
	介入戦略：極端	介入戦略：なだらか	意思決定のブレ：大	意思決定のブレ：小	交互作用：強	交互作用：弱
Naive	0.6335	0.8616	0.8650	0.4945	0.7286	0.8237
多クラス分類	0.6926	0.8713	0.8750	0.5970	0.7600	0.8549
条件付き独立	0.7015	0.8652	0.8722	0.6688	0.7536	0.8647
提案手法	0.7025	0.8728	0.8773	0.6621	0.7651	0.8661

多くの状況において、**提案手法は対応能力が高い**

ご清聴ありがとうございました

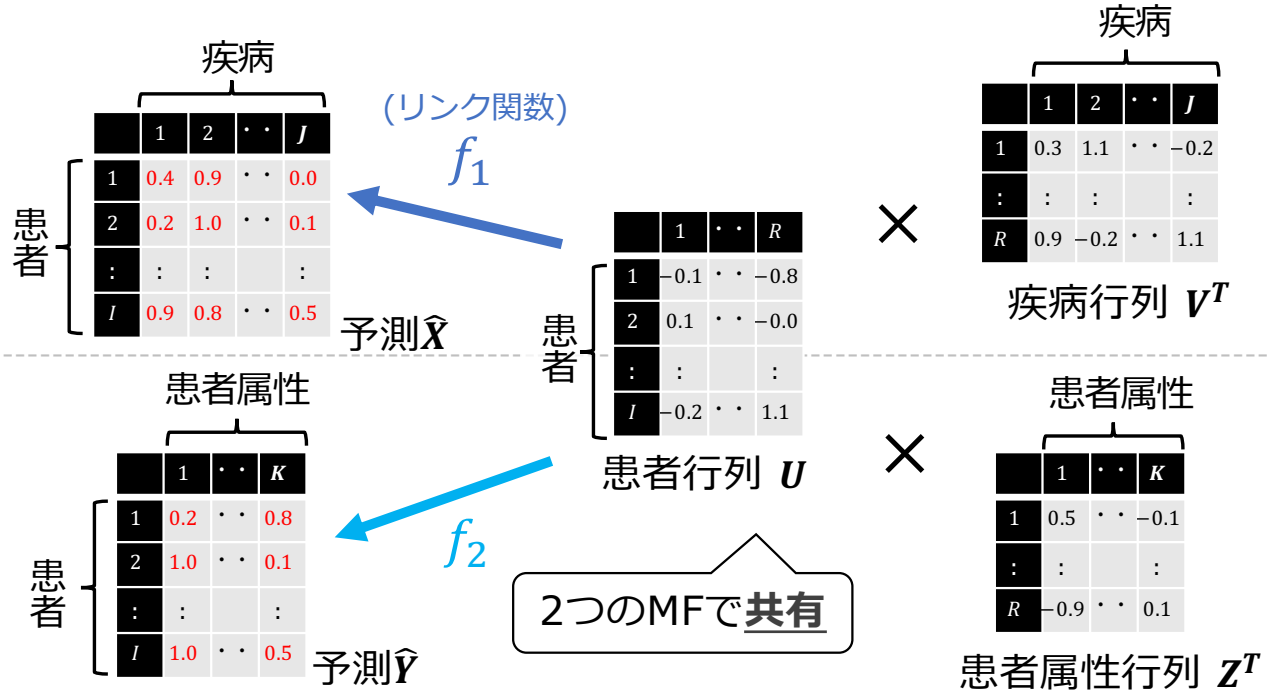
-Appendix-

※本来は発表内容でしたが、時間の都合上Appendixに置いたページがございます

CMF（Collective Matrix Factorization）[Singh, et al. (2008)]

（前提：患者の特徴量を考慮するため、**患者-患者属性の関係データ** $Y \in R^{I \times K}$ を導入）

- 2つの行列を同時に分解し、3つの行列を得る
- それぞれのMFは、リンク関数によって非線形変換を行う



メリット
- 複数の関係性の加味が可能
  (∵ 患者属性を特徴量として扱える)
- リンク関数による柔軟な出力

デメリット
- 特徴表現の解釈性が低い
  (∵ 要素に正と負の値が出現する)

NMF, CMFの比較

性質		NMF	CMF
解釈性	非負性	○	×
拡張性	複数の関係性の加味	×	○
	出力の柔軟性	×	○

目標の達成のためには、
「NMFの解釈性」と「CMFの拡張性」の両方が必要

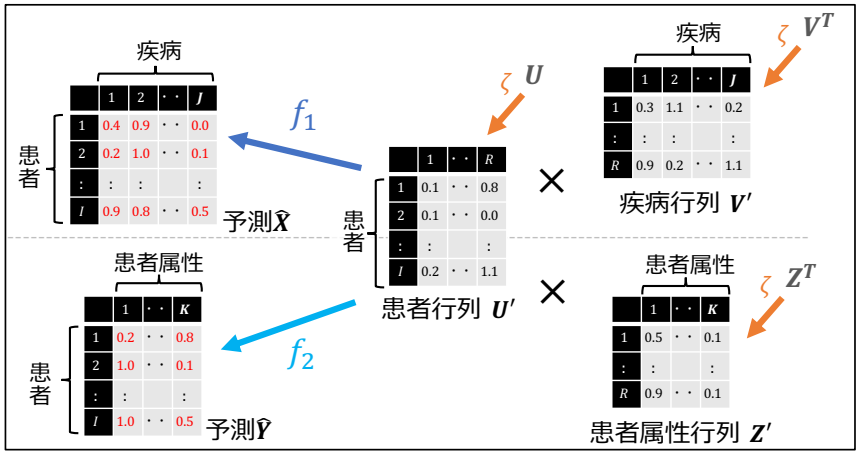
➡ 新たな手法を開発

PCMF（Positive Collective Matrix Factorization）

- 学習方法
  - 誤差逆伝播法を使用
    - 「連鎖律による**勾配計算**」＋「**最適化アルゴリズム**によるパラメータ更新」
  - $\hat{X} \cong X,\ \hat{Y} \cong Y$ となるように、パラメータ U, V, Z を更新する

- リンク関数
- 損失関数
- パラメータ初期値

設定方法は、論文に記載



Appendix

PCMF補足①

- リンク関数
  - 定義域が正であることを前提にしつつ、予測対象の性質・目的に応じ選択
  - 特に本研究の数値実験においては、両行列ともにシグモイド関数を適用

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

  - ただし、入力に対して一律に共通の正の数(パラメータ)を引く

$$\hat{\boldsymbol{X}} = \sigma(\boldsymbol{U}'\boldsymbol{V}'^{T} - \boldsymbol{C}_X)$$

$(\boldsymbol{C}_X \in \mathbb{R}^{I \times J}$：すべての要素が $c_X\ (> 0)$ の行列$)$