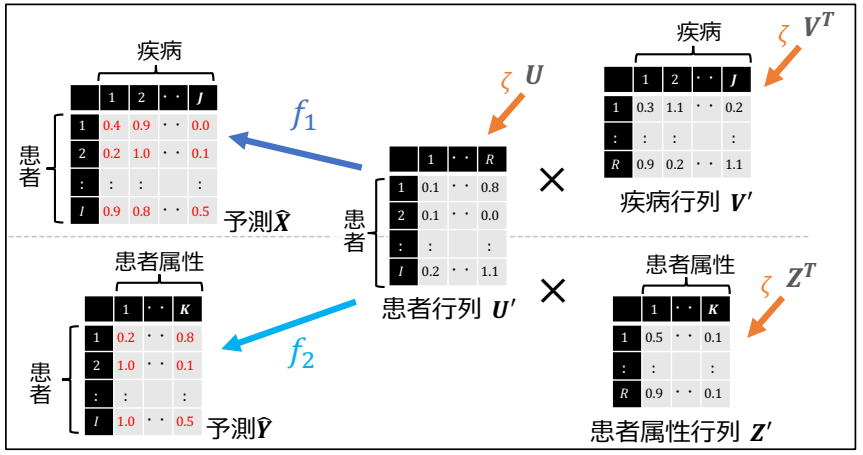
Appendix

PCMF補足②

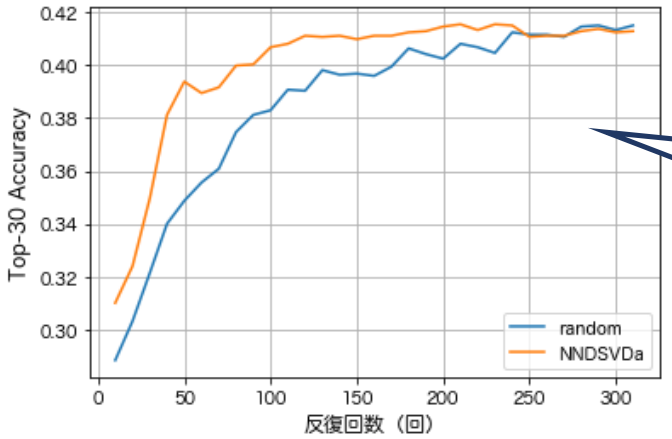
- 損失関数
  - 全体 $$L(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}) = D_X(\boldsymbol{X}, \hat{\boldsymbol{X}}) + \alpha D_Y(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) + \lambda\{\Phi_p(\boldsymbol{U}') + \Phi_p(\boldsymbol{V}') + \Phi_p(\boldsymbol{Z}')\}$$
  - 行列の近さ $$D_X(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \sum_{i,j} W^X_{i,j} \cdot d_X(X_{i,j}, \hat{X}_{i,j})$$
  - 要素の近さ $$d_X(x, \hat{x}) = -\{x \log \hat{x} + (1 - x) \log(1 - \hat{x})\}$$
  - 正則化項 $$\Phi_p(\boldsymbol{w}) = \|\boldsymbol{w}\|_p^p = \sum_i |w_i|^p$$



Appendix

PCMF補足③

- NNDSVDa [Christos, et al. (2008)] による初期化
  - NMFにおいて、元の行列にSVDを施し、特異ベクトルの負の成分を取り除きながら、2つの行列の初期値として構成させる手法
  - 特に、値が0となった要素を元の行列の全要素の平均値とする
  - PCMFにおいては、NNDSVDaによって得られた行列に、ソフトプラス関数の逆関数を用いて変換し、それぞれの行列の初期値として適用
  - 共有している行列については、それぞれの関係データから得られた初期値の加重平均



正規乱数で決定する方法(random)よりも、
NNDSVDaを用いる方法の方が収束が早い

Appendix

NNDSVDアルゴリズム

Step 1. Compute SVD: $\mathbf{X} = \mathbf{U\Sigma V}$

Step 2. Initialize $\mathbf{w}_1 = \mathbf{u}_1 \times \sqrt{\sigma_{1,1}}$ and $\mathbf{h}_1 = \mathbf{v}_1 \times \sqrt{\sigma_{1,1}}$.

Step 3. for j from 2 until k

$x = \mathbf{u}_j, y = \mathbf{v}_j,$

$px = pos(x), py = pos(y), nx = neg(x), ny = neg(y),$

$pn = norm(xp) \times norm(yp),\ nn = norm(nx) \times norm(ny),$

if $pn > nn,\ u = px/norm(px),\ v = py/norm(py),\ sigma = pn.$

else $u = nx/norm(nx),\ v = ny/norm(ny),\ sigma = nn.$

$\mathbf{w}_j = u \times \sqrt{\sigma_{j,j}}$ and $\mathbf{h}_j = v \times \sqrt{\sigma_{j,j}}$.

Step 4. Set all zeros of $\mathbf{W}$ and $\mathbf{H}$ equal to the average of all elements of $\mathbf{X}$.

$pos(x)/neg(x)$ returns vector with only positive/negative elements and 0. $norm(x)$ returns L_1-norm.

引用：朴玄信, 滝口哲也, and 有木康雄. "制約付き非負行列因子分解を用いた音声特徴抽出の検討." 情報処理学会研究報告音声言語情報処理 *(SLP)* 2008.123 (2008-SLP-074) (2008): 43-48.

Appendix

各要素で値の大きかった疾病

要素	ICD-10	疾病名	値
1	G530	帯状疱疹後神経痛	9.313
	B022	帯状疱疹，その他の神経系合併症を伴うもの	9.273
	G470	睡眠の導入及び維持の障害［不眠症］	9.252
	C251	膵体部	9.036
	R522	その他の慢性疼痛	9.010
2	J450	アレルギー性喘息を主とする疾患	10.386
	J459	喘息，詳細不明	8.840
	R620	標準発達遅延	8.829
	J209	急性気管支炎，詳細不明	8.769
	J304	アレルギー性鼻炎＜鼻アレルギー＞，詳細不明	8.721
3	H522	乱視	14.465
	H353	黄斑及び後極の変性	10.641
	H405	その他の眼疾患に続発する緑内障	10.263
	H250	老人性初発白内障	10.175
	H330	網膜剥離，網膜裂孔を伴うもの	10.131
4	I714	腹部大動脈瘤，破裂の記載がないもの	9.531
	C61	前立腺の悪性新生物＜腫瘍＞	9.470
	I48	心房細動及び粗動	8.975
	N40	前立腺肥大（症）	8.874
	C679	膀胱，部位不明	8.639

Appendix

各要素で値の大きかった疾病

5	O990	妊娠，分娩及び産じょく＜褥＞に合併する貧血	9.099
	O300	双胎妊娠	8.964
	P071b	その他の低出産体重（児）	8.729
	O908	産じょく＜褥＞のその他の合併症，他に分類されないもの	8.608
	O800	自然頭位分娩	8.605
6	K519	潰瘍性大腸炎，詳細不明	9.815
	K510	潰瘍性（慢性）全大腸炎	9.426
	M329	全身性エリテマトーデス＜紅斑性狼瘡＞＜ SLE ＞，詳細不明	8.749
	M1300	多発性関節炎，詳細不明	8.704
	B181	慢性 B 型ウイルス性肝炎，デルタ因子（重複感染）を伴わないもの	8.545
7	M932	離断性骨軟骨炎	7.661
	G442	緊張性頭痛	7.599
	R42	めまい＜眩暈＞感及びよろめき感	7.542
	M2399	膝内障，詳細不明	7.495
	M2441	関節の反復性脱臼及び亜脱臼	7.415
8	H71	中耳真珠腫	9.198
	H959	耳及び乳様突起の処置後障害，詳細不明	8.817
	H902	伝音難聴，詳細不明	8.549
	C220	肝細胞癌	8.449
	H740	鼓室硬化症	8.379

Appendix

各要素で値の大きかった疾病

9	J690	食物及び吐物による肺臓炎	8.830
	K918	消化器系のその他の処置後障害，他に分類されないもの	8.202
	C169	胃，部位不明	8.103
	C162	胃体部	7.993
	R090	窒息	7.861
10	C029	舌，部位不明	8.738
	C031	下顎歯肉	8.361
	C443	その他及び部位不明の顔面の皮膚	8.262
	S141	頚髄のその他及び詳細不明の損傷	7.873
	C770	頭部，顔面及び頚部リンパ節	7.822
11	I456	早期興奮症候群	11.087
	I490	心室細動及び粗動	9.807
	I471	上室（性）頻拍（症）	9.349
	I472	心室（性）頻拍（症）	9.171
	I420	拡張型心筋症	9.128
12	C549	子宮体部，部位不明	8.757
	D383	縦隔	7.953
	C73	甲状腺の悪性新生物＜腫瘍＞	7.925
	D150	胸腺	7.795
	C509	乳房，部位不明	7.703

疾病の特徴表現解析

● 疾病同士の類似性解析

E11と同時にリスクとなりやすい疾病

■ **E11**(2型糖尿病)と類似度の高い疾病を抽出

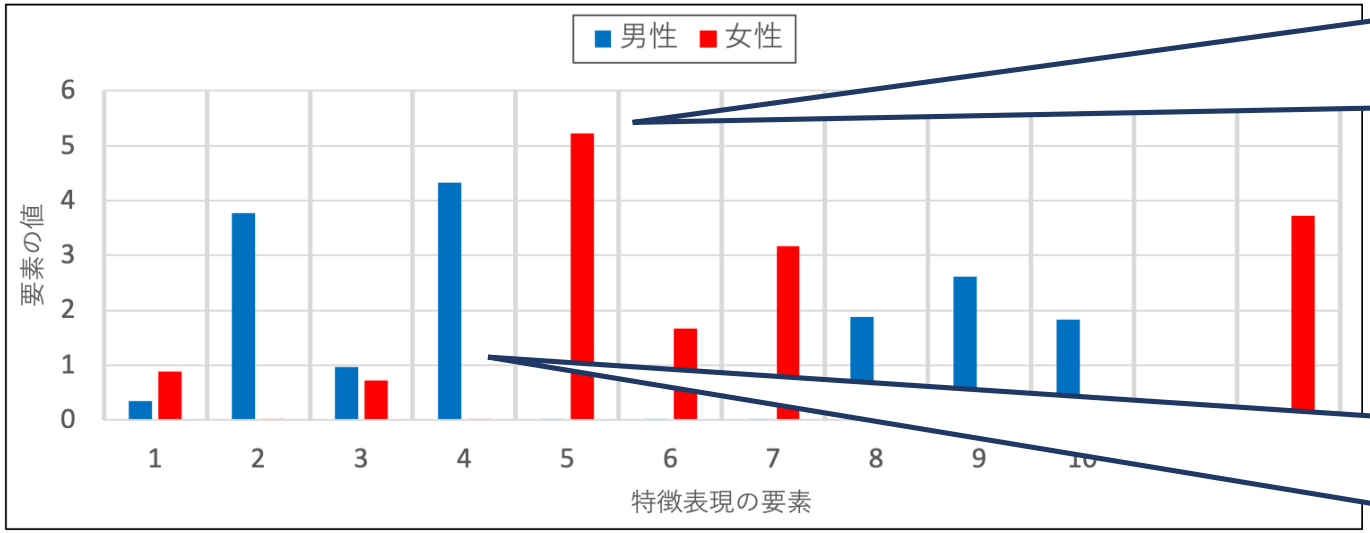
類似度	ICD-10	疾病名	E11発症率※
0.941	E14	詳細不明の糖尿病	15.1%
0.912	I67.2	脳動脈のアテローム粥状硬化(症)	26.3%
0.904	E14.2	詳細不明の糖尿病，腎合併症を伴うもの	25.8%
0.900	E78.0	純型高コレステロール血症	23.4%
0.896	G59.0	糖尿病性単ニューロパチー	38.9%
0.888	N08.3	糖尿病における糸球体障害	25.3%

※全体でのE11発症率：5.24%

あらゆる疾病同士の**関連性・類似性**も把握できる

患者属性の特徴表現解析

■ 例：患者属性「男性」「女性」の特徴表現



ICD-10	疾病名
O99.0	妊娠，分娩及び産褥に合併する貧血
O30.0	双胎妊娠
P07.1b	その他の低出産体重（児）
O90.8	産褥のその他の合併症
080.0	自然頭位分娩

ICD-10	疾病名
I71.4	腹部大動脈瘤，破裂の記載がないもの
C61	前立腺の悪性新生物＜腫瘍＞
I48	心房細動及び粗動
N40	前立腺肥大（症）
C679	膀胱，部位不明

因子の意味解析から、**患者・疾病・患者属性のもつ特徴**を解釈できる

Appendix

疾病×患者属性の特徴表現解析

- 疾病の患者属性の類似性解析

表 5.7. 「男性」と類似度の高い疾病

ICD-10	疾病名	類似度	「男性」率 (%)	「男性」の中の発症率 (%)	(参考) 全体発症率 (%)
N433	精巣＜睾丸＞水瘤，詳細不明	0.872	100.0	0.6	0.3
C61	前立腺の悪性新生物＜腫瘍＞	0.830	99.3	5.6	2.8
K409	一側性又は患側不明の鼠径ヘルニア，閉塞及び壊疽を伴わないもの	0.825	79.4	2.2	1.4
J040	急性喉頭炎	0.812	56.7	0.4	0.4
T782	アナフィラキシーショック，詳細不明	0.807	50.7	0.5	0.4
N209	尿路結石，詳細不明	0.799	69.8	0.4	0.3
J439	肺気腫，詳細不明	0.793	87.6	1.9	1.1
J439	肺気腫，詳細不明	0.793	87.6	1.9	1.1
N200	腎結石	0.789	63.2	0.6	0.4
M1099	痛風，詳細不明	0.787	81.8	1.9	1.1
R798	その他の明示された血液化学的異常所見	0.783	58.6	1.8	1.6

Appendix

患者の特徴表現解析

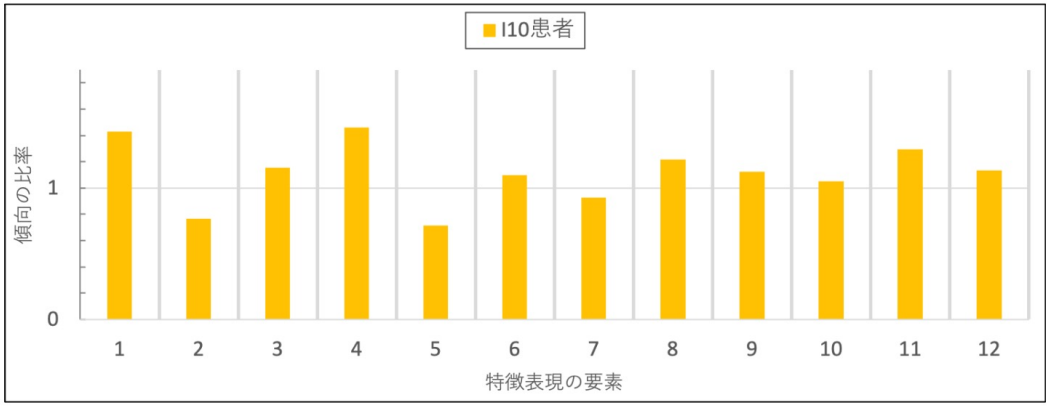


図 5.5. I10 患者の特徴表現の傾向

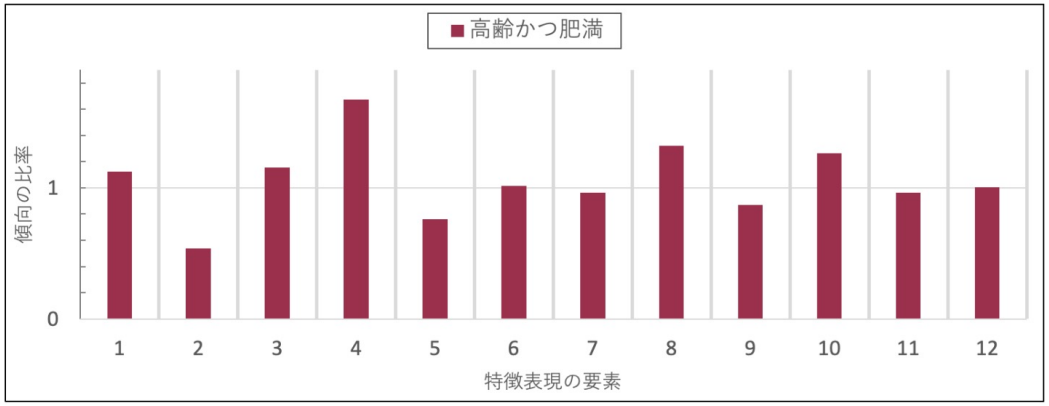


図 5.6. 「高齢かつ肥満」患者の特徴表現の傾向

Appendix

提案手法の貪欲アルゴリズム

Algorithm 1 貪欲アルゴリズム

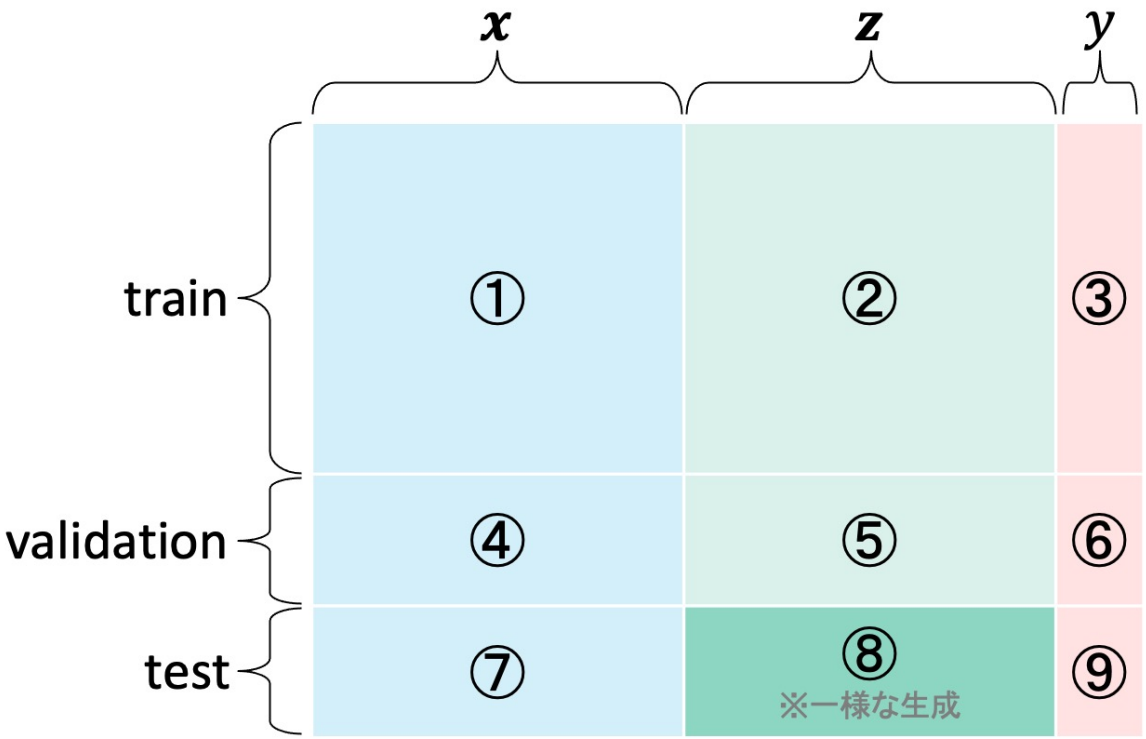
Input: $p(z_k|\mathbf{x})$ $(k = 1, 2, \cdots, K)$，およびそれらの予測精度（AUC など）．改善値の下限 Δ.

Output: 介入 z_k $(k = 1, 2, \cdots, K)$ 間の因果構造

while 最大の改善値 $\geq \Delta$ **do**
  for $k = 1, 2, \cdots, K$ **do**
    for $k' = 1, 2, \cdots, K$ $(k' \neq k)$ **do**
      if 因果関係 $z_{k'} \rightarrow z_k$ が元々存在せず，加えても有向非巡回でない． **then**
        z_k を目的とする予測に $z_{k'}$ を一時的に条件（説明変数）として加え，z_k の予測精度およびその改善値を求める．
      end if
    end for
  end for
  最大の改善値を得られる因果関係について，Δ 以上の場合に認め，因果構造に加える．
end while

Appendix

データの扱い方



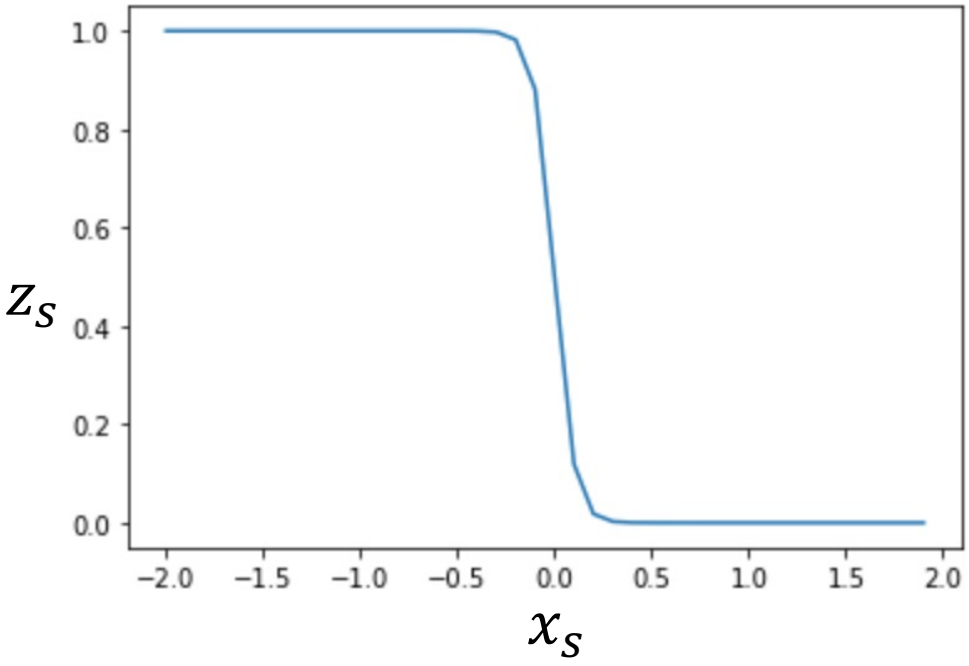
- ①-②で$p(z|x)$の学習
- ①②-③で回帰モデルの学習
- ④-⑤で$p(z|x)$の検証
  - 特に因果探索で使う
- ④⑤-⑥で回帰モデルの検証
  - IPW lossで計算
- ⑦⑧-⑨で回帰モデルのテスト

Appendix

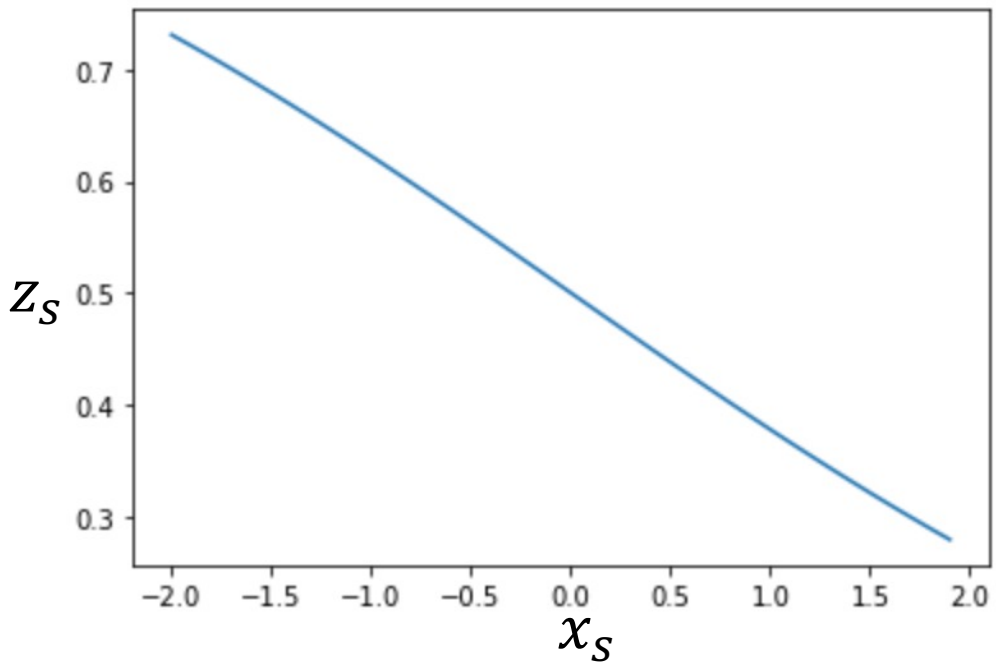
パラメータによるデータの傾向変化

- 介入戦略の強さ（α）

$\alpha = 20$



$\alpha = 0.5$

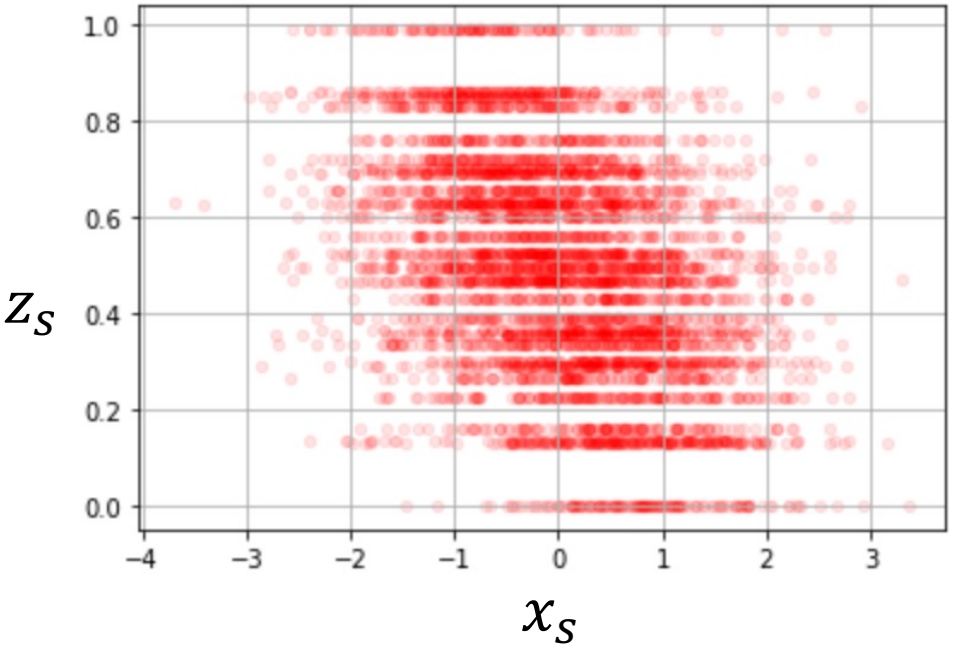


Appendix

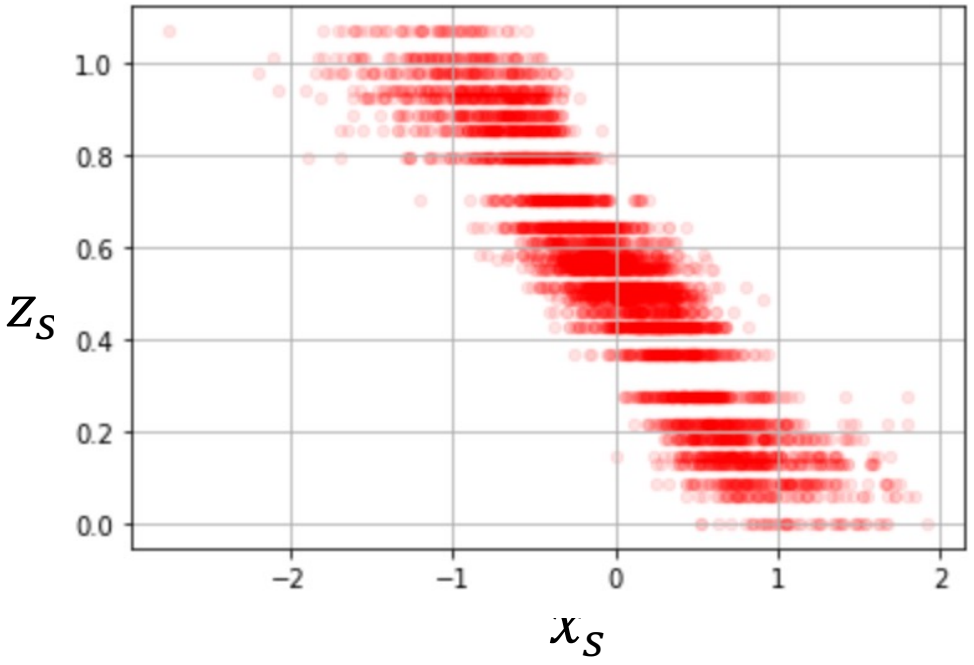
パラメータによるデータの傾向変化

- 意思決定のブレの大きさ（σ^2）

$\sigma^2 = 0.16$



$\sigma^2 = 0.01$

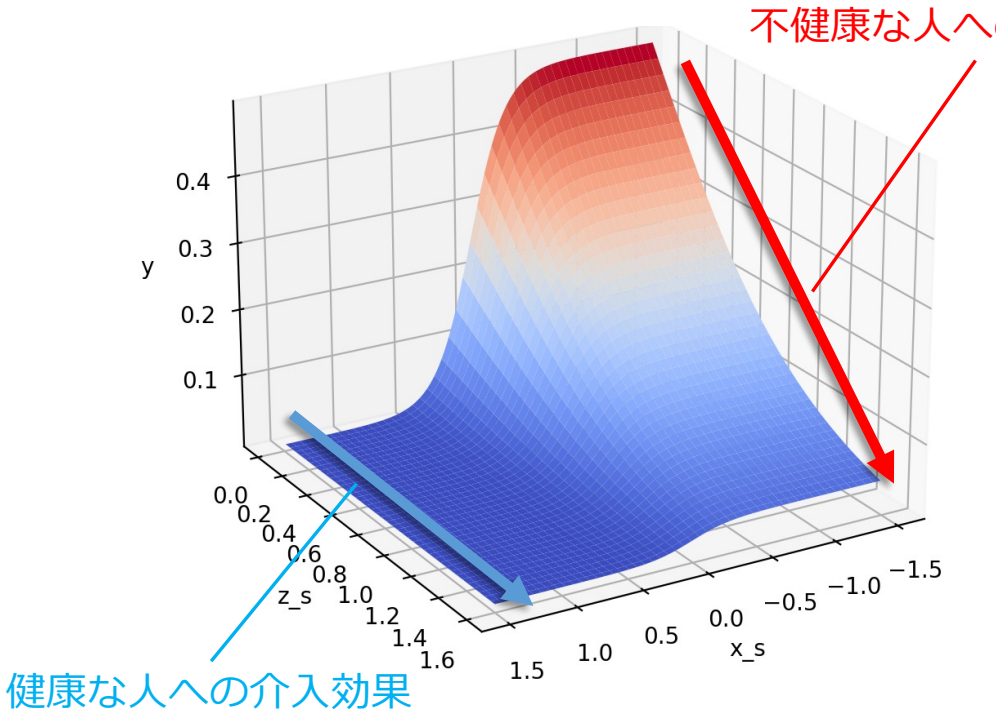


Appendix

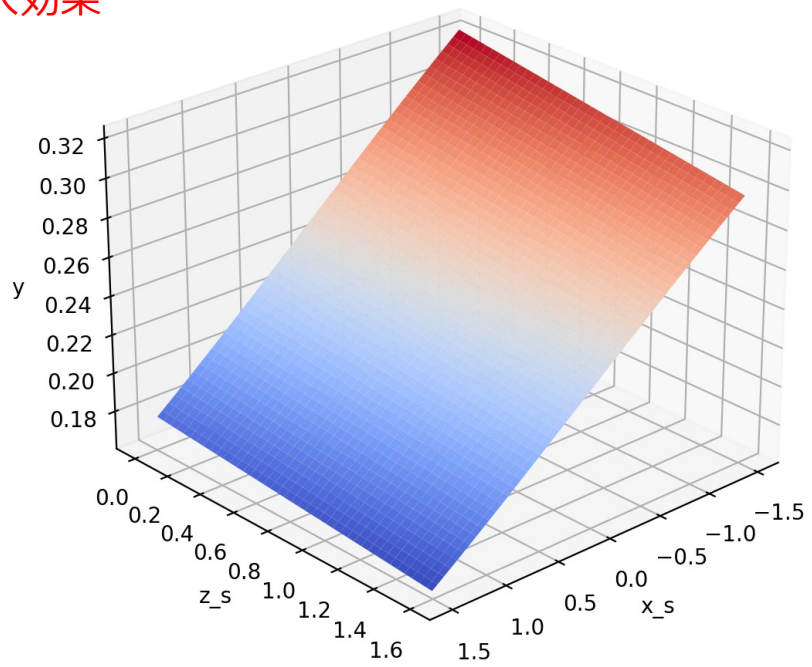
パラメータによるデータの傾向変化

- 交互作用の大きさ（$[\beta_1, \beta_2]$）

$[\beta_1 = 7, \beta_2 = 2]$



$[\beta_1 = 0.4, \beta_2 = 0.1]$



Appendix

モデルのパラメータ

（記載のないものはscikit-learnのデフォルト）

- ロジスティック回帰
  - 'max_iter':100
- Random Forest
  - 'max_depth': 10, 'max_features': 0.8, 'min_samples_leaf': 6, 'n_estimators’: 100
- 因果探索
  - AUCの最低改善値：0.01