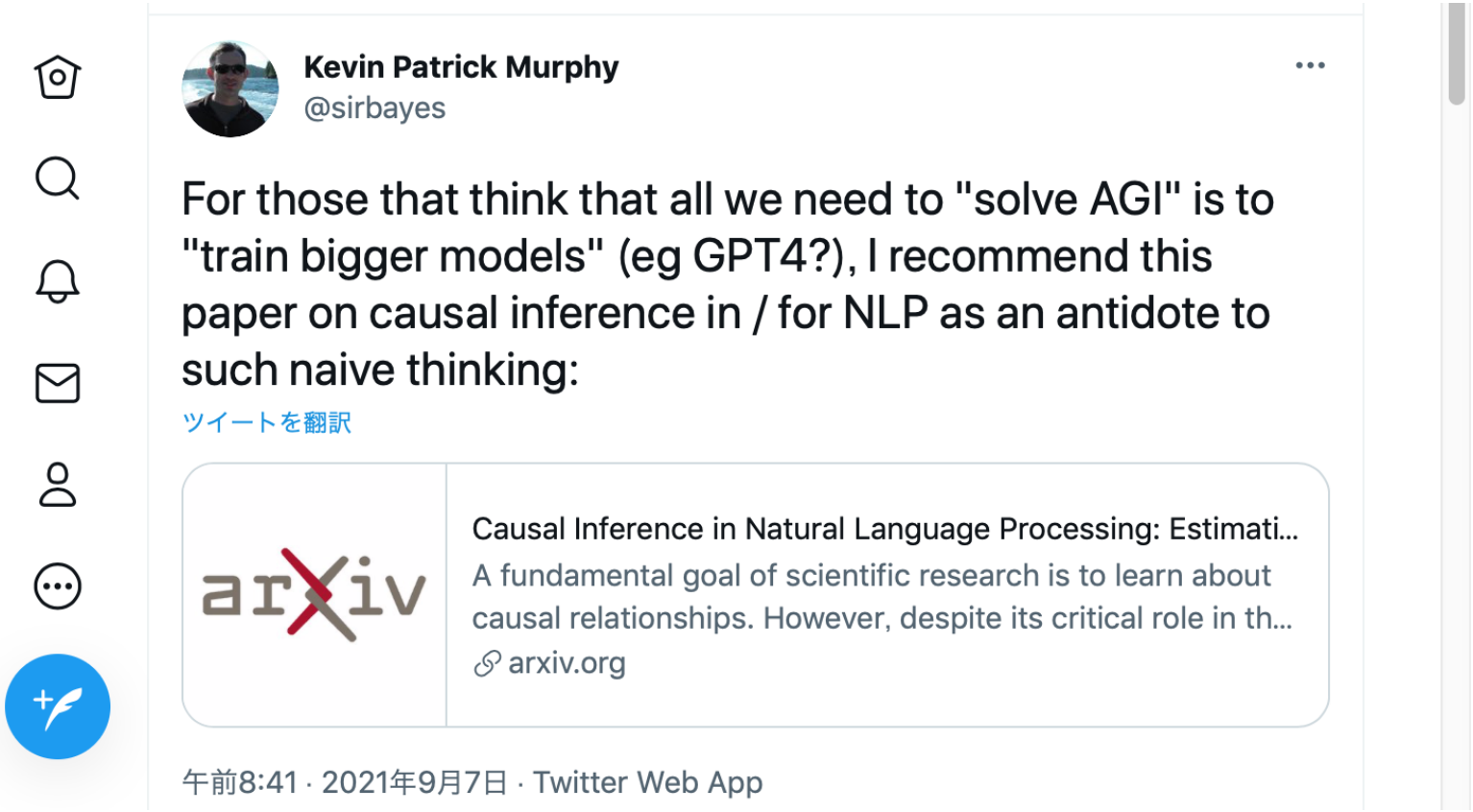
“Adapting Text Embeddings for Causal Inference” (Veitch, Sridhar, Blei: UAI 2019)

持橋大地 (統計数理研究所)
daichi@ism.ac.jp
最先端NLP 2021
2021-9-16 (木)

つい先日...



- Kevin Murphy (MLaPPの著者)のツイート (9月7日)

Causal Inference for NLP

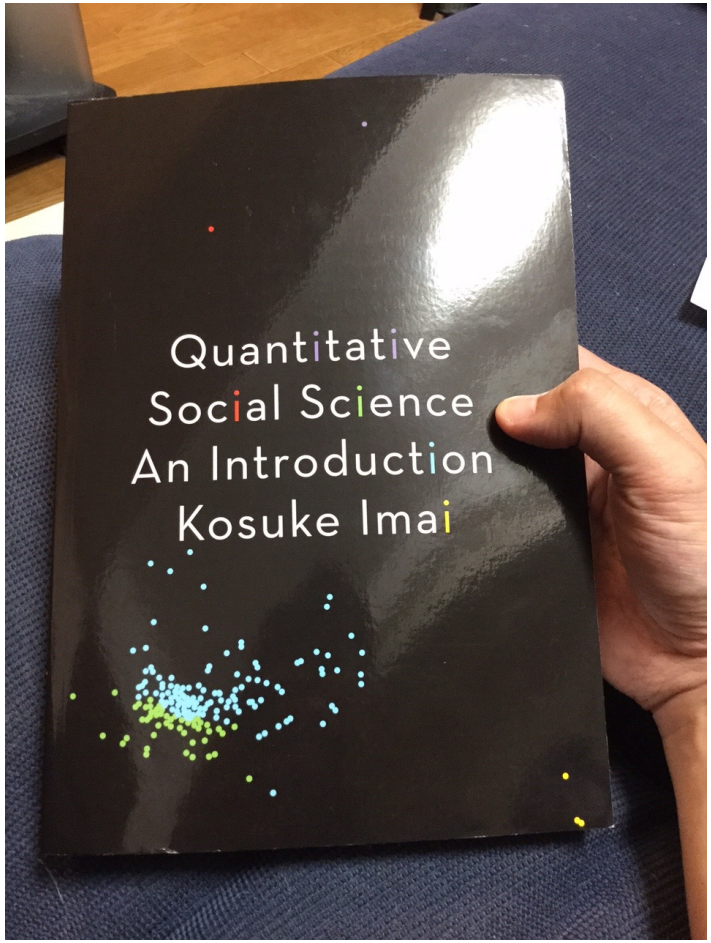
Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond

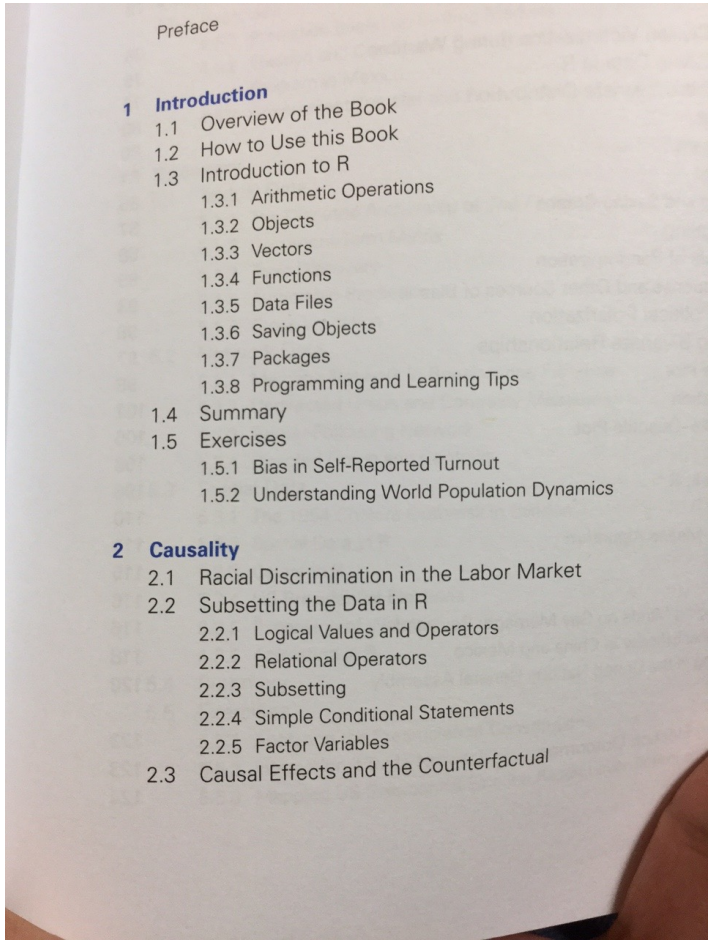
Amir Feder[1], Katherine A. Keith[2], Emaad Manzoor[3], Reid Pryzant[4], Dhanya Sridhar[5], Zach Wood-Doughty[6], Jacob Eisenstein[7], Justin Grimmer[4], Roi Reichart[1], Margaret E. Roberts[8], Brandon M. Stewart[9], Victor Veitch[7,10], and Diyi Yang[11]

[1]Technion - Israel Institute of Technology
[2]University of Massachusetts Amherst
[3]University of Wisconsin - Madison
[4]Stanford University
[5]Columbia University
[6]Johns Hopkins University
[7]Google Research
[8]University of California San Diego
[9]Princeton University
[10]University of Chicago
[11]Georgia Tech

- Feder+ (2021) のサーベイ論文 (arXiv: 2109.00725v1 cs.CL)
  - その他、多くの有名人の共著

社会科学では、因果推論は基本的





- Imai (2016)では、2章がいきなり因果推論の話題

どんな論文?

Adapting Text Embeddings for Causal Inference

Victor Veitch* **Dhanya Sridhar*** **David M. Blei**

Department of Statistics and Department of Computer Science
Columbia University

- “Causally sufficient embedding” を提案
  - BERTベース(C-BERT)とLDAベース(C-ATM)の両方
- 文書をそのまま使うよりロバストな推論が可能
- 実験により、真の因果関係をより正確に復元できた

因果推論とは

私も初心者なのでおこがましいですが…

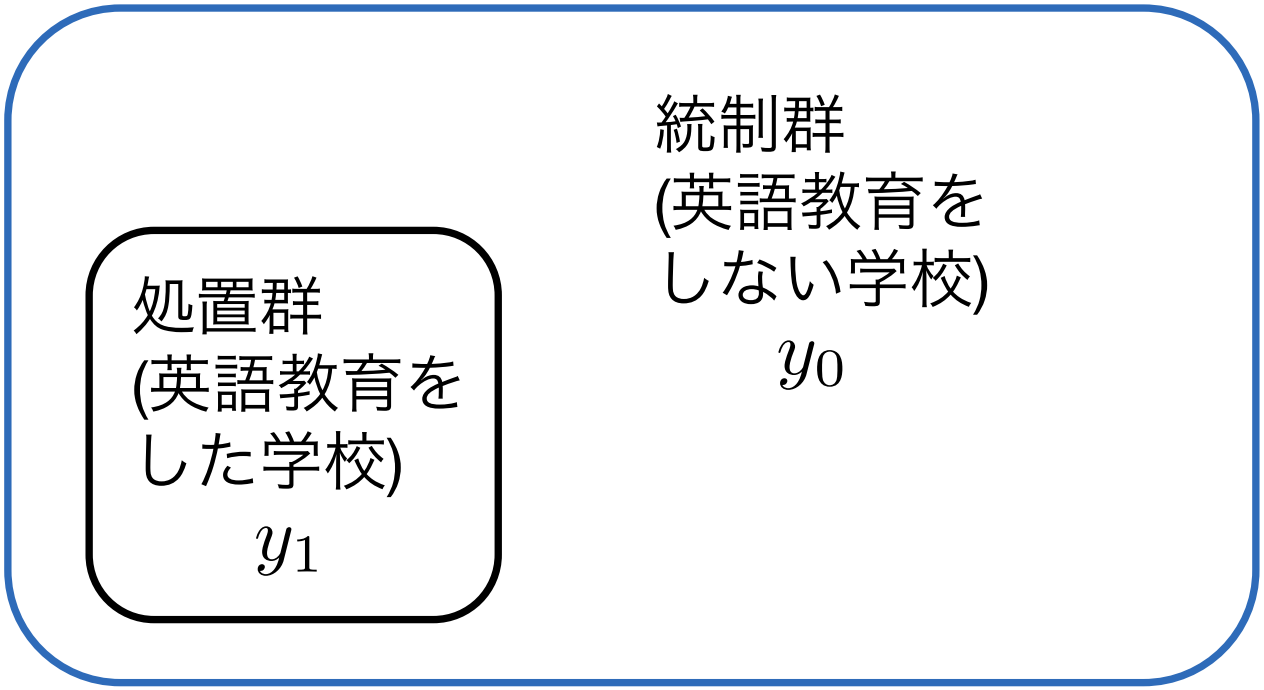
- 小学校に英語教育を導入することの効果を知りたい
  → 中学生の終わりに、英語力をテストする
- 「英語教育を導入した小学校の卒業生」の点数と、「英語教育を導入しなかった小学校の卒業生」の点数を単に比較すればよいか?

  もちろん×！
  - 英語教育を導入した小学校は、もともと教育熱心な学校で、子供の知能も平均的に高い
  - 英語教育を導入しなかった小学校は、荒れている可能性もあり、子供の知能ももともと高くない

  可能性が高い
  これを選択バイアスという

処置群、統制群と選択バイアス



- 処置群 (treatment) は、それ以外 (control) と同じ分布からランダムに選ばれているわけではない！
  - ランダムに選ばれている状況がRCT (A/Bテストなど)

因果推論とは (2)

- どうすればフェアに比較できる？

  小学校の特徴量を導入する (共変量 x：covariate)
  - 地域の親の大卒者の割合
  - 地域の親の平均収入
  - 生徒の中で帰国子女が占める割合
  - ...
- x の情報を使って、観測された英語力を補正する
- どうやって行うか?

因果推論の分野では、
confounderと言うことも
多い

因果推論とは (3)

- 因果推論の理論は色々とあるが、結局、次の2つが基本的：
  (1) 回帰モデルを使う方法
  (2) 傾向スコアを用いる方法

- 共通する考え方
  「もしtreatmentがあった場合の結果」
  ー「もしtreatmentがなかった場合の結果」
  の期待値を計算する (Rubinの反実仮想モデル)
  - 各小学校について、「英語教育を行った場合の得点」
    ー「英語教育を行わなかった場合の得点」
    の期待値を求める

回帰モデルを使う方法

- 共変量を *x* (例: 小学校の特徴量)、
  効果を知りたい量を *y* (例: 中学校での英語テストの得点)、
  とおくと、*z*=1/0が小学校で英語教育をしたかどうか
  を表すとき、

$$
\begin{aligned}
E[y_1 - y_0] &= E_x[y_1 - y_0 \mid x] \\
&= E_x\left[E[y_1|x] - E[y_0|x] \mid x\right] \\
&= E_x\left[E[y_1|x, z=1] - E[y_0|x, z=0] \mid x\right]
\end{aligned}
$$

共変量がわかれば、
割り当てに関わらず
yの期待値は同じ

- よって、*x*→*y*を予測する回帰モデルを直接学習して、
  それを使って差の期待値を計算すればよい (論文 (2.3)式)

$$
\hat{\psi}^Q = \frac{1}{\sum_i t_i} \sum_i t_i \left[\hat{Q}(1, z_i) - \hat{Q}(0, z_i)\right]. \qquad (2.3)
$$

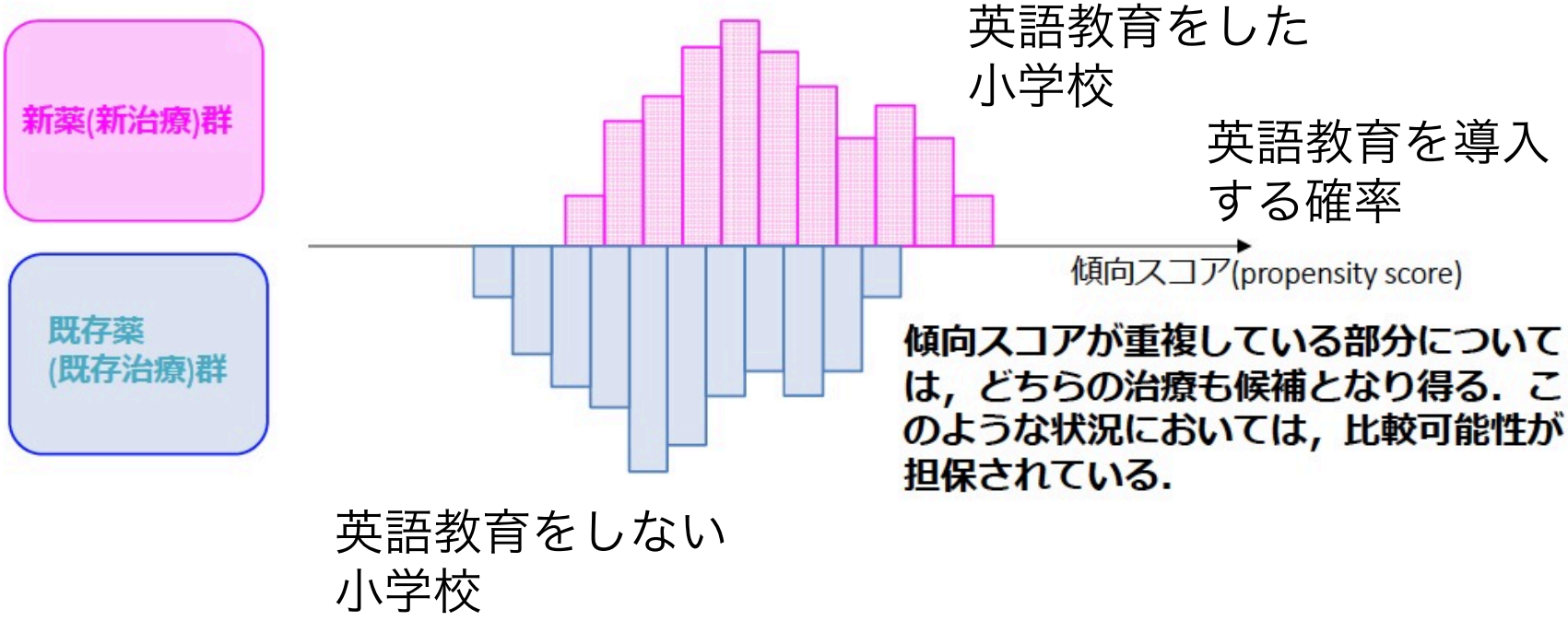
傾向スコアを用いる方法

- 傾向スコア (propensity score)：共変量*x*の下で、各ユニットが処置群に割り当てられる確率
  - 特徴量*x*を持つ小学校で、英語教育が導入される確率
  - propensityというより、tendencyというと分かりやすい
- 式で書くと、

$$g(x) = p(T = 1|x)$$

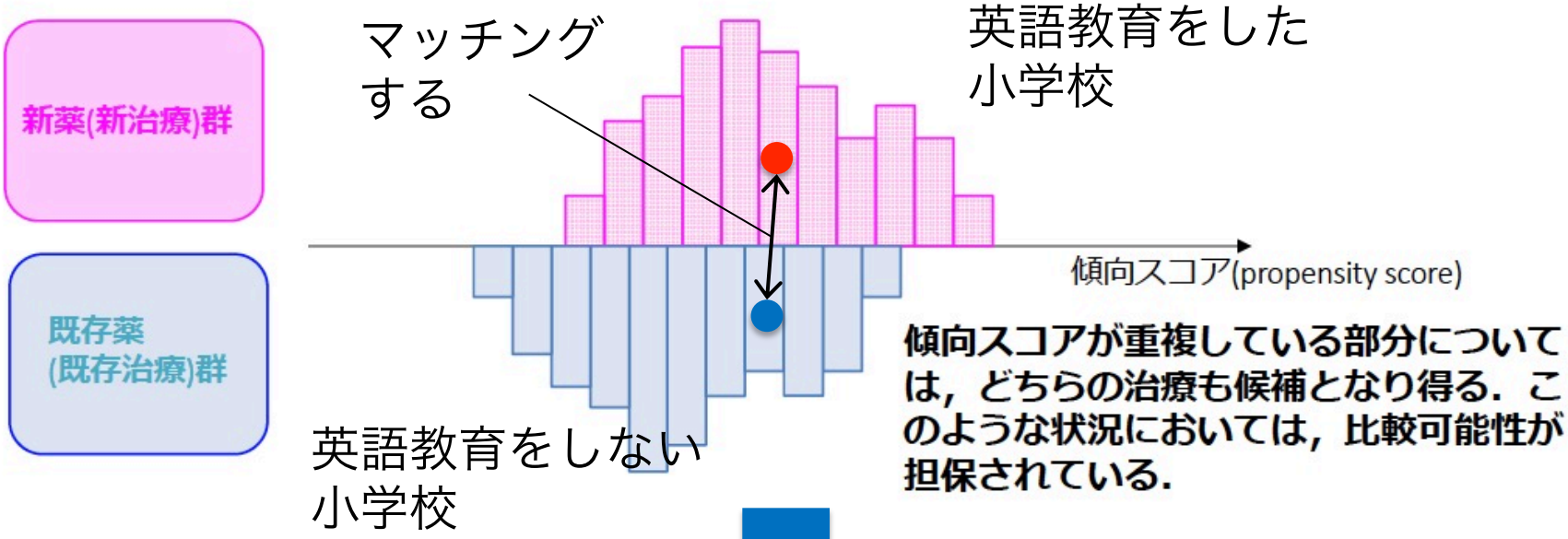
  論文では x の代わりにzを用いている
- この傾向スコアがあると、小学校の特徴による違いを吸収できる → 傾向スコアが同じ小学校を比べれば、フェアな比較結果が得られる

傾向スコアを用いる方法

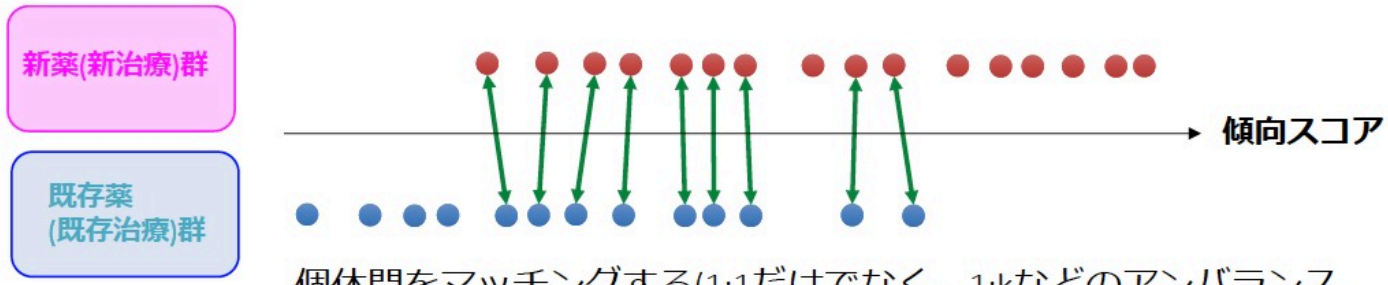


下川先生 (和歌山県立医大)のスライドより引用

傾向スコアを用いる方法 (2)

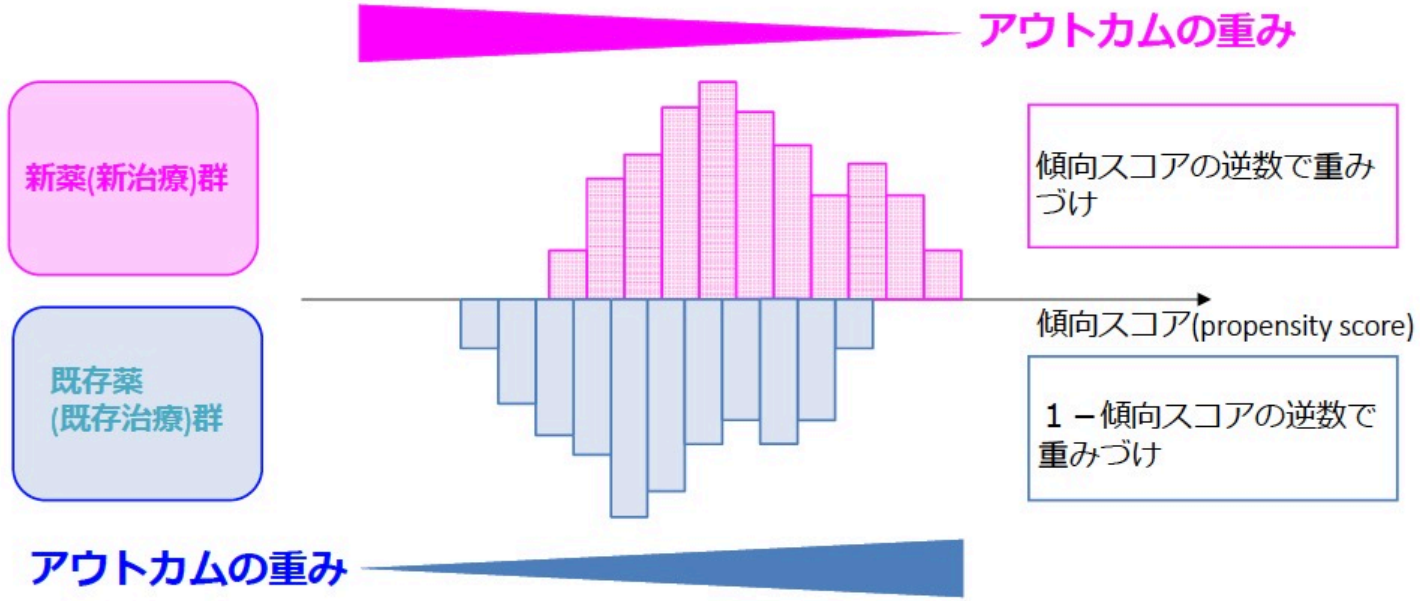


■ ペアマッチング

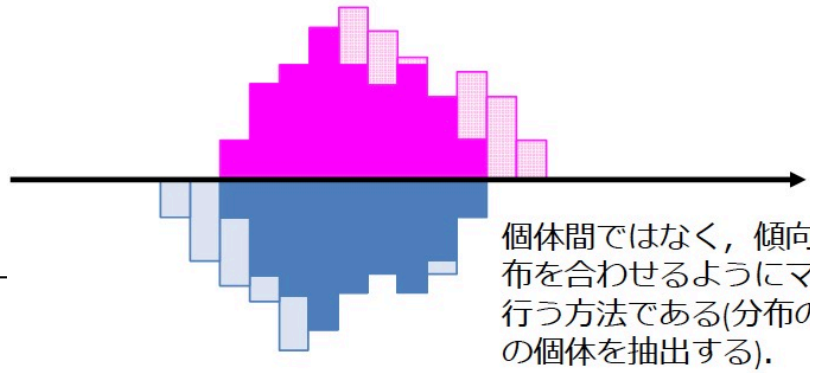


傾向スコアを用いる方法 (3)

- 別の方法：重み付けを行うことで、2つの分布を近づける (IPW: Inverse Probability Weighting)



- 論文では(2.2)式



結局..

(1) 回帰モデルを使う場合、共変量*x*→結果*y*の回帰関数を学習

(2) 傾向スコアを使う場合、上の回帰関数と、傾向スコア(確率)を出力する関数の両方を学習

- テキストでは、単語列wをそのまま使うのは悪手
  → 埋め込み λ(w) を学習する
  - 学習するのは、λ(w)→*y* の回帰関数と、
    λ(w)→z の傾向スコアを求める関数
  - (λ(w), λ(w)→y, λ(w)→z) のtripleを学習する

因果埋め込みの学習

- BERTの場合 (Causal BERT)

$$
\begin{aligned}
L(\mathbf{w}_i;\xi,\gamma) =& \big(y_i - \tilde{Q}(t_i,\lambda_i;\gamma)\big)^2 && \leftarrow \text{出力の予測}\\
&+ \mathsf{CrossEnt}\big(t_i,\tilde{g}(\lambda_i;\gamma)\big) && \leftarrow \text{割り当ての予測}\\
&+ L_{\mathrm{U}}(\mathbf{w}_i;\xi,\gamma). && \leftarrow \text{言語モデル}
\end{aligned}
$$

- LDAの場合 (Causal ATM, Amortized Topic Model)

$$
\begin{aligned}
L(\mathbf{w}_i;\eta,\beta,\gamma) =& -\mathcal{L}_i(\beta,\eta) && \leftarrow \text{出力の予測}\\
&+ \mathbb{E}_{q(\theta|\mathbf{w};\eta)}\big[\mathsf{CrossEnt}\big(t_i,\tilde{g}(\theta_i;\gamma)\big] && \leftarrow \text{割当予測}\\
&+ \mathbb{E}_{q(\theta|\mathbf{w};\eta)}\big[y_i - \tilde{Q}(t_i,\theta_i;\gamma)^2\big]. && \leftarrow \text{言語モデル}
\end{aligned}
$$

実験：NLPでの因果推論

- 論文に定理が含まれていると、Acceptされやすくなるのか?
  - 定理が含まれるような論文は技術レベルが高いので、そもそも採択されやすい可能性
  - データ: PeerRead arXiv論文採択データ, 2891/11778本
  - 仮想レスポンス: $Y_i \sim \text{Bernoulli}(\sigma(0.25t_i + b_1(\pi(\tilde{z}_i) - 0.2)))$
- SNSのポスト(Reddit)に「女性」タグが付いていると、いいねの数が少なくなるのか?
  - 女性が選ぶ話題や書き方によって、反応が変わっている可能性
  - データ: Redditの3つのフォーラム, 90k個
  - 仮想レスポンス: $Y_i = t_i + b_1(\pi(\tilde{z}_i) - 0.5) + \varepsilon_i \quad \varepsilon_i \sim N(0, \gamma).$

実験結果

- 言語モデルと教師あり次元削減の効果
  (推定結果がGround truthに近いかどうか)

(a) Language Modeling Helps

Dataset:	Reddit (NDE)	PeerRead (ATT)
Ground truth	1.00	0.06
Unadjusted	1.24	0.14
NN $\hat{\psi}^{Q}$	1.17	0.10
NN $\hat{\psi}^{\text{plugin}}$	1.17	0.10
BERT (sup. only) $\hat{\psi}^{Q}$	0.93	0.19
BERT (sup. only) $\hat{\psi}^{\text{plugin}}$	1.17	0.18
C-ATM $\hat{\psi}^{Q}$	1.16	0.10
C-ATM $\hat{\psi}^{\text{plugin}}$	1.13	0.10
C-BERT $\hat{\psi}^{Q}$	1.07	0.07
C-BERT $\hat{\psi}^{\text{plugin}}$	1.15	0.09

(b) Supervision Helps

Dataset:	Reddit (NDE)	PeerRead (ATT)
Ground truth	1.00	0.06
Unadjusted	1.24	0.14
BOW $\hat{\psi}^{Q}$	1.17	0.13
BOW $\hat{\psi}^{\text{plugin}}$	1.18	0.14
BERT $\hat{\psi}^{Q}$	-15.0	-0.25
BERT $\hat{\psi}^{\text{plugin}}$	-14.1	-0.28
LDA $\hat{\psi}^{Q}$	1.20	0.07
LDA $\hat{\psi}^{\text{plugin}}$	1.20	0.09
ATM $\hat{\psi}^{Q}$	1.17	0.08
ATM $\hat{\psi}^{\text{plugin}}$	1.17	0.08

実験結果 (2)

- 論文データでのATT

Confounding:	Low	Med.	High
Ground truth	0.06	0.05	0.03
Unadjusted	0.08	0.15	0.16
NN $\hat{\psi}^{Q}$	0.05	0.10	0.30
NN $\hat{\psi}^{\text{plugin}}$	0.05	0.10	0.30
C-ATM $\hat{\psi}^{Q}$	0.07	0.10	0.32
C-ATM $\hat{\psi}^{\text{plugin}}$	0.07	0.10	0.32
C-BERT $\hat{\psi}^{Q}$	0.09	0.07	0.04
C-BERT $\hat{\psi}^{\text{plugin}}$	0.10	0.09	0.05

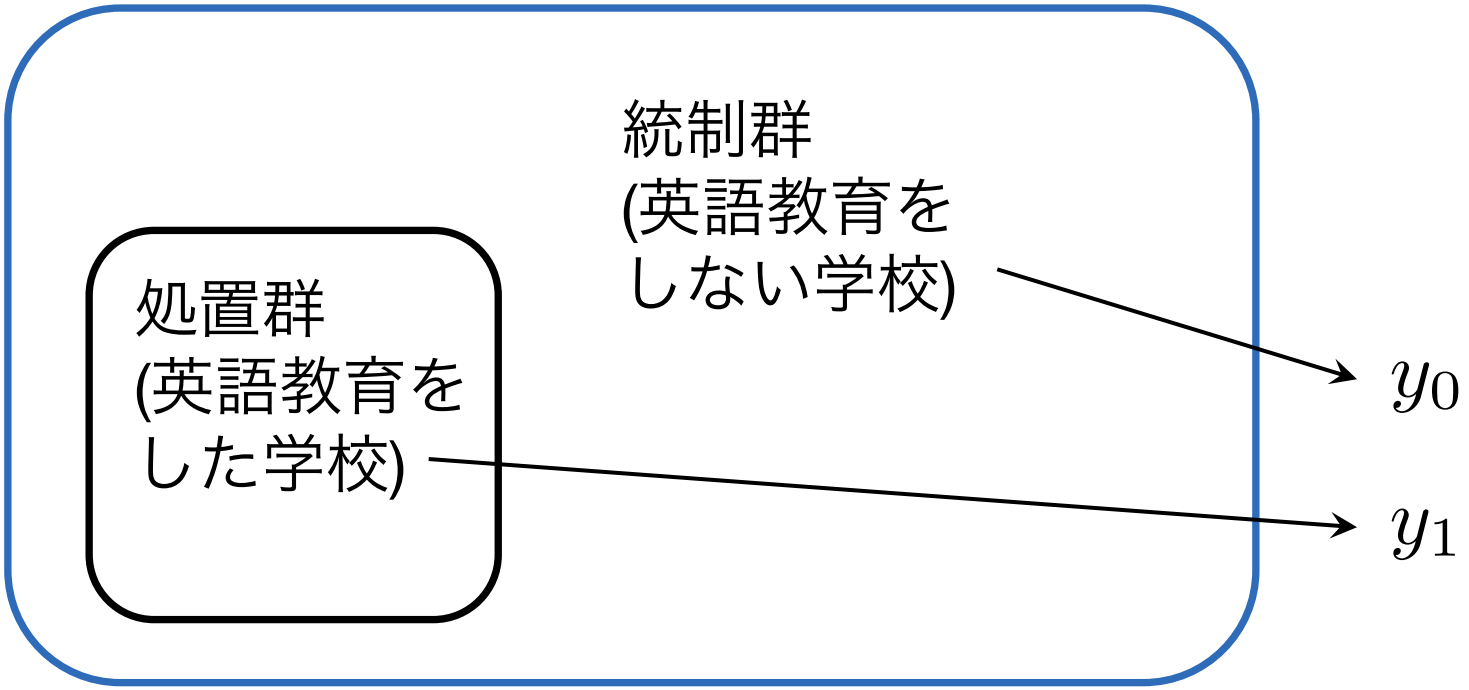
- Redditでの性別の効果

	okcupid	childfree	keto
Unadjusted	-0.18 ± 0.01	-0.19 ± 0.01	-0.00 ± 0.00
C-BERT $\hat{\psi}^{Q}$	-0.10 ± 0.04	-0.10 ± 0.04	-0.03 ± 0.02
C-BERT $\hat{\psi}^{\text{plugin}}$	-0.15 ± 0.05	-0.16 ± 0.05	-0.01 ± 0.00

結論と展望

- テキストの因果効果を埋め込みに縮約するような、causally sufficient embedding を提案している
  - 目的関数に言語モデル項も含まれている(あった方が性能が良い) → 数学的には何の意味なのか？
- 将来の課題
  - 埋め込みが何を意味しているか、可視化したりテストして検証できる必要がある
  - 処置と出力がどちらもテキストの外にあるが、テキスト自体が含まれる場合 (eg. 女性だと文体がどう変わるか?) もある
  - 分野適応に近い方法なので、因果推論の手法を使って分野適応を改善できるか？

処置群と統制群のイメージ (再掲)



- 現状の分野適応は、直接、高次元のyへの回帰関数を学習してしまっている (バイアス大)
  → 処置群になる1次元の傾向スコアを経由した方がいいのでは？

メッセージ

- 医学や社会科学、ビジネス(広告効果測定など) だけでなく、NLPでも因果推論は必要かつ有効
- 共変量 (covariate, confounder) が超高次元なので、適切な埋め込みを用いて対処する必要
  → causally sufficient embedding
- 言語モデルの効果など、まだ理論面はこれからという印象
- 他のモデルとの接続？