FY2025/2025年度
Causal Inference Team
Shohei Shimizu
因果推論チーム　清水 昌平

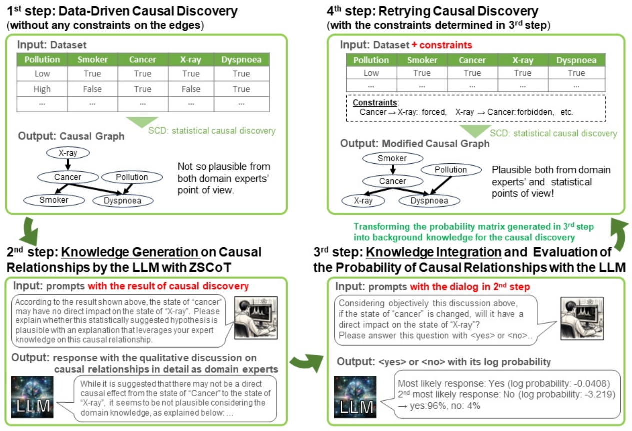
RIKEN
AIP Center for Advanced Intelligence Project

Causal Discovery
因果探索

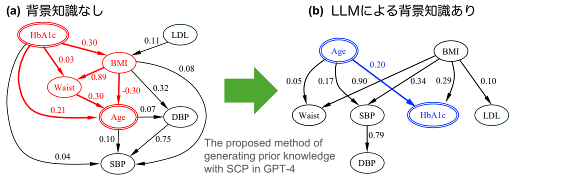
■ LLMから得られた背景知識や制約を、因果探索に統合

M. Takayama, T. Okuda, T. Pham, T. Ikenoue, S. Fukuma, S. Shimizu, A. Sannai.
Integrating large language models in causal discovery: a statistical causal approach.
Transactions on Machine Learning Research (TMLR), 2025.

■ Statistical Causal Promptingを提案



(a) 背景知識なし　(b) LLMによる背景知識あり



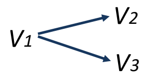
データが少ないと失敗

リークのない (LLMが知らない)
健康診断データで評価

Modeling Causal Processes
因果プロセスの圏論的モデル化

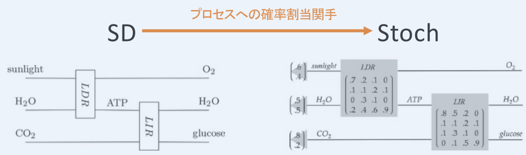
Otsuka, J., Hayashi, T., Yoshii, T., Saigo, H. (2025).
Modeling Causal Processes, *Synthese*. 206.

DAGモデルの因果
＝出来事間の関係



因果のプロセスとしての側面をモデル化できないか？

プロセスの圏（SD）から確率行列の圏（stoch）への関手functorとしての因果モデル

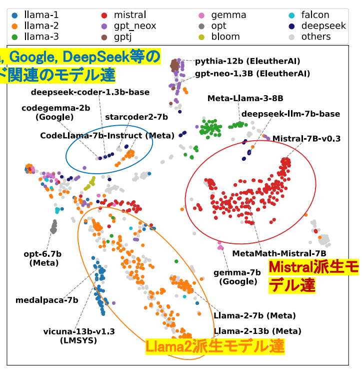


プロセス的モデリングの利点

1. マルコフ条件を前提としない（共通原因で条件付けても、結果同士は従属しうる）
2. 異なる粒度のモデル（関手）間の同一性の定式化
→ 因果的抽出 causal abstractionへの理論的定式
（Otsuka & Saigo 2022, *PMLR* ）

Map of Large Language Models
大規模言語モデルの地図

自然言語処理分野のトップ会議 ACL 2025 において Outstanding Paper Awradを受賞



モデルタイプ(ベースモデル)

・Hugging Faceには約30万個の言語モデルが公開されていて，モデル比較や性能評価は重要な課題

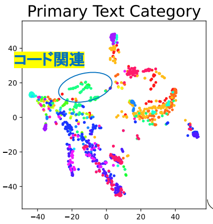
・多数のテキストに対する言語モデルの対数尤度ベクトルをモデル座標とする方法を提案

・実際に1018個の言語モデルにおいて，Pileコーパスからサンプリングした10000個のテキストに対する対数尤度ベクトルを使用して地図を作成

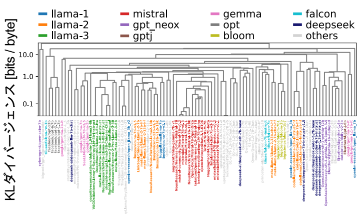
・モデル地図のユークリッド距離の２乗がKLダイバージェンスの近似となる定理 (下平 1993)が基礎

・客観的にモデルの個性（特性）を定量化した可視化だけでなく，ベンチマーク性能を高精度に予測できる

Primary Text Category



得意なテキスト分野（尤度で測定）　テキスト分野名



Momose Oyama, Hiroaki Yamagiwa, Yusuke Takase, Hidetoshi Shimodaira (arXiv:2502.16173, ACL2025 main), Mapping 1,000+ Language Models via the Log-Likelihood Vector

Factor analysis 因子分析

Terada, Y. (2026+) Statistical properties of matrix decomposition factor analysis, conditionally accepted for *Journal of Computational and Graphical Statistics*.

Factor model: $x = \Lambda f + \epsilon \ \rightleftharpoons\ \Sigma := \mathrm{Var}(x) = \Lambda\Lambda^\top + \Psi^2$

- x: p-dimensional observation, $\Lambda \in \mathbb{R}^{p\times m}$: factor loading matrix with $m < p$
- f : m-dim random vector with $\mathbb{E}[f] = 0_m$, $\mathrm{Var}(f) = I_m$, and $\mathbb{E}[f\epsilon^\top] = O_{m\times p}$
- ϵ : p-dim random vector with $\mathbb{E}[\epsilon] = 0_p$ and $\mathrm{Var}(\epsilon) = \Psi^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$

Matrix Decomposition Factor Analysis (MDFA)

For $\Phi := [\Lambda, \Psi]$ and $Z := [F, E]$ $(F \in \mathbb{R}^{n\times m},\ E \in \mathbb{R}^{n\times p})$,　F, E : nuisance params of size $O(n)$

$$\mathcal{L}_n(\Phi, Z) = \frac{1}{n}\left\|X_n - (F\Lambda^T + E\Psi)\right\|_F^2 = \frac{1}{n}\sum_{i=1}^{n}\|x_i - (\Lambda f_i + \Psi e_i)\|^2,$$

where $1_n^T F = 0_m^\top$, $1_n^\top E = 0_p^\top$, $F^\top F/n = I_m$, $E^\top E/n = I_p$, and $F^\top E = O_{m\times p}$.

Pros: Stable optimization via ALS :), **Cons:** No theoretical guarantees :(

Q: Is MDFA truly "factor analysis"? This has been an open problem for nearly **20 years**.

⇒ MDFA = Minimum Discrepancy Function estimator with the Wasserstein distance!

Proposition (Concentrated loss or Profile likelihood). *For any* $\Phi \in \Theta_\Phi$ $(\Sigma(\Phi) := \Phi\Phi^\top)$,

$$\mathcal{L}_n(\Phi) = \min_{Z\in\Theta_Z}\mathcal{L}_n(\Phi, Z) = \mathrm{tr}(\widehat{S}_n) + \mathrm{tr}\{\Sigma(\Phi)\} - 2\mathrm{tr}\left\{\left(\widehat{S}_n^{\frac{1}{2}}\Sigma(\Phi)\widehat{S}_n^{\frac{1}{2}}\right)^{\frac{1}{2}}\right\} = d_{\mathrm{BW}}^2\left(\widehat{S}_n, \Sigma(\Phi)\right),$$

where $d_{\mathrm{BW}}(A, B)$ *is* **the Bures-Wasserstein distance** *between p.s.d. matrices* A *and* B.

Consistency and asymptotic normality of the MDFA estimator follow directly from this fact.