High-Dimensional Statistical Modeling Team
Makoto Yamada
高次元統計モデリングチーム　山田 誠

RIKEN Center for Advanced Intelligence Project

高次元統計モデリングチームの研究概要

目標

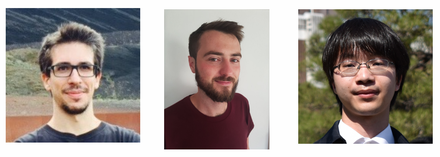
- 医療, 材料分野において**科学的発見をサポートする機械学習基盤**の構築
  - 重要な特徴をデータから精度よく容易に見つけられる手法の研究開発
  - 機械学習研究者以外でも容易に利用可能なソフトウェア開発
- 新規の科学的発見を容易にする基盤を確立し, **ヘルスケア, 材料, 農業等の分野で革新**を目指す
  - 医療費の削減 (個別化医療, 疾病予測)
  - 材料発見の大幅な効率化

By Guillaume Paumier, from https://commons.wikimedia.org/wiki/File:DNA_microarray.svg

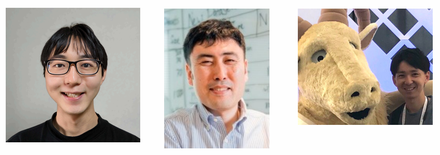
2021年度主要成果

- **カーネル選択的推論　（重要な特徴の選択）**
  - HSIC Lassoの選択的推論 (ICML 2021)
  - Knockoff filterに基づいた選択的推論 (AISTATS 2022)
- **最適輸送（高次元データの類似度計算）**
  - 教師あり木構造Wasserstein距離 (ICML 2021)
  - 不均衡木構造最適輸送の研究 (AISTATS 2021)
  - NASへの応用　(NeurIPS2021)
  - 異ドメイン間の木構造最適輸送 (AISTATS 2021)
  - 木構造Barycenter推定 (AISTATS 2022)

- **高次元・構造データの学習**
  - ヘシアン行列近似手法の提案 (arXiv 2021)
  - Lassoの高速スクリーニング法 (NeurIPS 2021)
  - ランダム特徴量に基づいた学習 (SDM 2021)

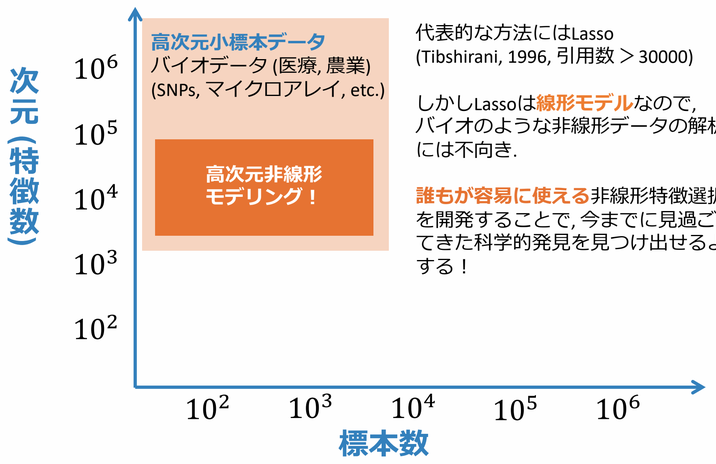


- **機械学習応用**
  - 医療画像データからの学習



主要成果1:高次元非線形特徴選択

高次元小標本データ



代表的な方法にはLasso
(Tibshirani, 1996, 引用数 > 30000)

しかしLassoは**線形モデル**なので,
バイオのような非線形データの解析
には不向き.

誰もが容易に使える非線形特徴選択手法を開発することで, 今までに見過ごされてきた科学的発見を見つけ出せるようにする！

解釈性

- 少数特徴で高い予測性能
- 特徴の信頼度がわかる　（仮説検定）

チャレンジ

- 非線形モデルは複雑　　（誰も利用しない）
- 非凸最適化が利用される（最適化が難しい）

HSIC Lassoの選択的推論 (ICML 2021)


我々のチームでは2020年度までに**カーネル法に基づいた非線形選択的推論**の確立をした。

- 選択的独立性検定　(AISTATS 2018, 2020)
- 選択的二標本検定　(ICLR 2019)
- 選択的適合度検定　(NeruIPS 2019)

本年度は我々のチームで開発している**HSIC Lasso法の選択的推論**を提案しICML 2021にて報告。

Table 2: P-Values Of The Partial Target For The Single-cell RNAseq Dataset

Gene	p-value	Gene	p-value
ACTB	0.961	*IGJ*	**< 0.001**
CD14	**0.026**	*LYZ*	**< 0.001**
FCER1A	**< 0.001**	*MTRNR2L2*	0.420
FCGR3A	**0.001**	*RPS3A*	**< 0.001**
FTL	0.968	*TMSB4X*	**0.012**
HLA-DPA1	**< 0.001**	*TVAS5*	0.553
IFI30	**0.002**		

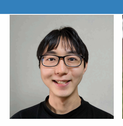
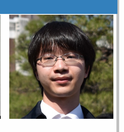
遺伝子発現量データから重要な特徴を選択したところCD14, FCGR3A (DC4), FCER1A (DC2 and DC3), FTL (DC4), IFI30 (cDC-like), IGJ (pDC-like), LYZ (cDC-like) は重要であることが報告されていた (Science 2017). **HLA-DPR1 は元論文では報告されておらず新規のバイオマーカーの可能性がある.**

カーネル選択的推論 (AISTATS 2022)


- Knockoff filterを初めてカーネル法(HSIC, MMD)に適用.
  - K個の特徴を選択
  - Knockoff filterを用いて検定を行う
- FDRのコントロールが可能
- Github
  PeterJackNaylor/knockoff-MMD-HSIC

主要成果2: 最適輸送

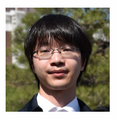
構造のある最適輸送

我々のチームでは最適輸送の研究に取り組んでおり特に**構造のある最適輸送**の研究に取り組んでいる

- 複数木に基づいた木構造Wasserstein距離 NeurIPS 2019
- 不均衡木構造最適輸送の研究 NeurIPS 2020, AISTATS 2021
- Neural Architecture Search (NAS)への応用 NeurIPS2021
- 異なるドメイン間の木構造最適輸送 AISTATS 2021
- 密度比推定に基づいた最適輸送 ECML 2021
- 木構造データにおけるBarycenter推定 arXiv 2019, 2021
- 高次元データのための最適輸送 arXiv 2020
- Wasserstein距離の再評価 arXiv 2021

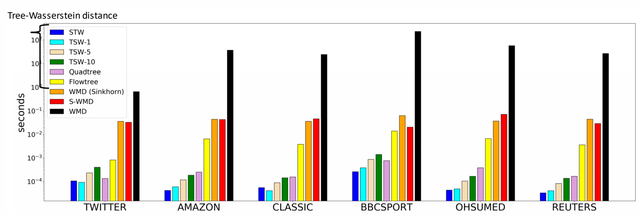
教師あり木構造 Wasserstein距離 (ICML 2021)



Wasserstein距離 (WD)は自然言語処理で盛んに利用されているが**計算時間が非常に大きい問題**がある。

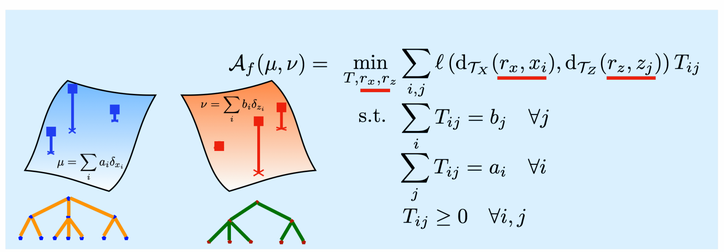
WDと同等の性能を得られつつ、WDの数百倍高速に計算可能な**教師あり木構造Wasserstein距離**を提案。

木構造Wasserstein距離: $W_{d_\mathcal{T}}(\mu_i,\mu_j)=\left\|\mathbf{w}_v\circ(\mathbf{I}-\mathbf{D}_{\text{par}})^{-1}\begin{pmatrix}\mathbf{0}_{|V_{\text{in}}|}\\ \mathbf{a}_i-\mathbf{a}_j\end{pmatrix}\right\|_1$

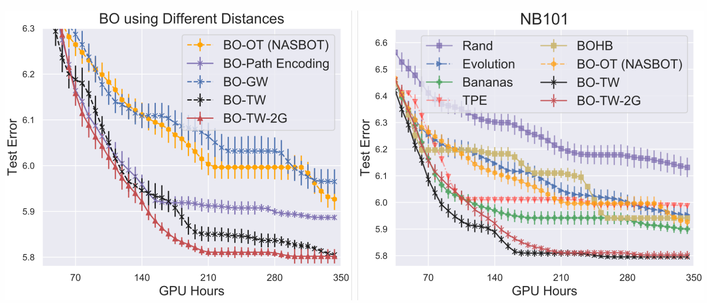


木構造Gromov-Wasserstein (AISTATS 2021)

- 超高速に異なる空間の距離を測る方法



Wasserstein 距離のNASへの応用(ICML 2021)



主要成果3:高次元・構造データの学習

GNNの限界と克服方法を示した (NeurIPS2019, SDM2021)

GNN は特定のグラフ対を**どんなパラメータを用いても**区別で きない、という問題が知られている。

$$\forall\Theta \quad \mathrm{GNN}_\Theta\left(\begin{array}{c}\diagup\\ \triangle\end{array}\right)=\mathrm{GNN}_\Theta\left(\square\right)$$

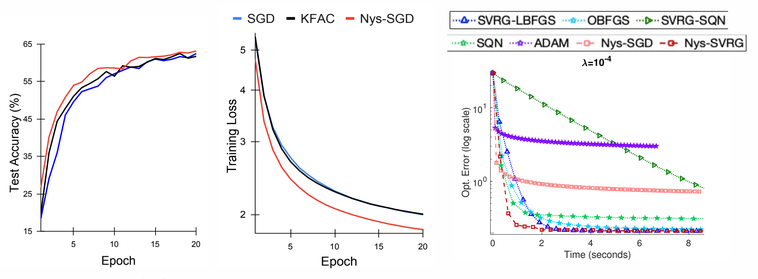
- GNNと局所分散アルゴリズムの類似性を指摘しGNNの近似度を示した。
- 局所分散アルゴリズムでは決定的なものよりも乱択アルゴリズムの方が良い近似度を達成できることを利用し**ノード特徴量に乱数を連結する方法を提案。**

グラフニューラルネットワーク (GNN) というモデルの限界を示し、その限界を克服する方法を示した。

Nystrom近似ニュートン法 (arXiv 2021)

- Nystrom近似に基づいたニュートン法

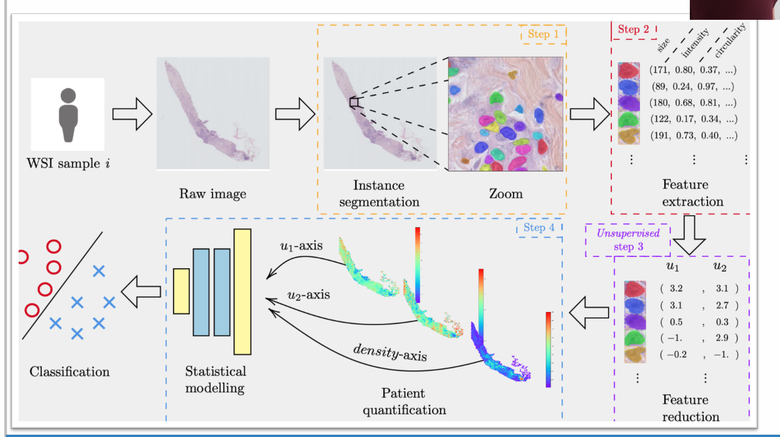
$$\boldsymbol{w}_t=\boldsymbol{w}_{t-1}-\eta(\boldsymbol{Z}_\tau\boldsymbol{Z}_\tau+\rho\boldsymbol{I})^{-1}\boldsymbol{v}_{t-1}$$



Non-convex (EfficentNet on ImageNet)　　Convex

主要成果4:機械学習応用

解釈可能な組織プロファイリングと分類



2022年度の目標

- 高次元データからの因果推論方法の提案
- 最適輸送技術の医療応用
- 新規バイオマーカーの発見