

高次元非線形統計モデリング

High-dimensional Nonlinear Statistical Modeling

山田 誠

理研革新知能統合研究センター

JST さきがけ研究員 (社会デザイン)

統計数理研究所, Aalto University (客員)



研究目標

Research Goal

- 医療, 材料, 農業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
 - 重要な素性・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
 - 非線形解析手法の研究開発
 - 科学的根拠となる指標(p値)のための理論構築を目指す
 - 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 (R, Python, Matlab等)
- 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起こす
 - 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

研究目的 (続き)

Research Goal (Continue)

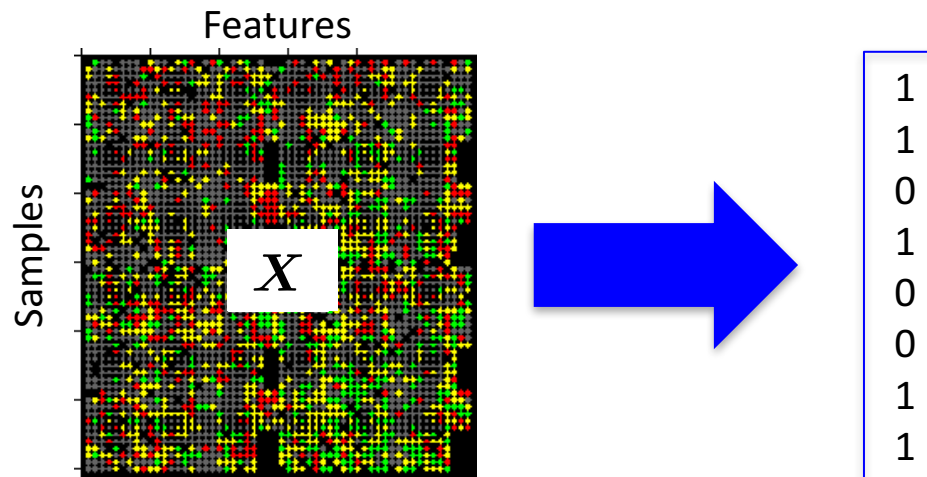
- 見過ごされていた事実を効率良く発見したい
 - 現在は線形手法といった, 比較的単純な方法を用いてデータ解析をしていることが多く, 重要な結果が見過ごされている可能性が高い.
 - 高度な手法 (例: 非線形手法) を用いることで, 線形方法では見つけられなかったことを発見したい!
 - 予測には線形手法で十分だが、解釈のためには非線形手法が重要といえる (かもしれない).
 - グラフデータなどの構造化されているデータからの科学的発見

医療、材料、農業分野の研究者とタッグを組み, 提案する枠組みを既存データに適用し, 新規の科学的発見を目指す.

特徴選択

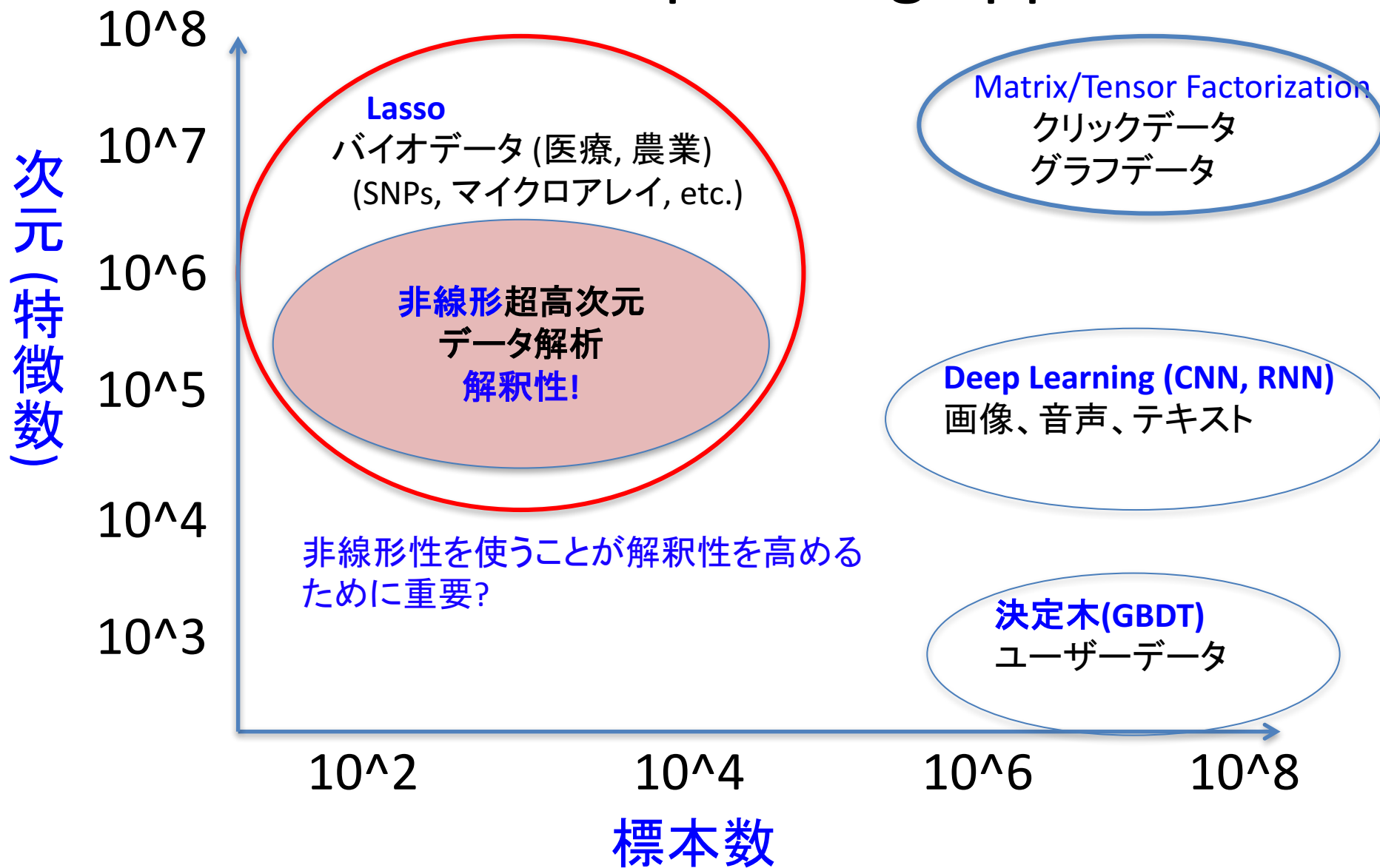
Feature Selection

- 入力・出力ベクトル: $\boldsymbol{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$
- 学習データ: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)$
- **ゴール**: d 次元の入力 $\boldsymbol{x}$ から, 出力 y に関連する r 個の特徴を選択する.
- **応用例**: バイオマーカー検出



データ 及び 関連手法

Data and corresponding approach



高次元非線形特徴選択

High-dimensional Nonlinear Feature Selection

- ゴール: 科学的発見のための機械学習方法の構築
- 解釈性 (a.k.a., 特徴選択?)
 - 少数特徴で高い予測性能
 - 高次元データにおいて線形手法は予測性能が高いが、多くの特徴が必要 (解釈性は低い)
 - p値
 - 各特徴のp値がわかると嬉しい
 - 個人の特徴を選択
- チャレンジ
 - 非線形モデルは複雑になりがち
 - よくわからないのであまり使われない?
 - 非凸最適化がよく利用される
 - サンプル数が少ないのですぐにオーバーフィットする.
- 単純&高性能(&凸最適化).

機械学習基盤

Machine Learning framework

- **非線形特徴選択**に基づく機械学習基盤の構築
 - 基礎非線形アルゴリズムの研究開発
 - ソフトウェア開発



機械学習 × サイエンス

Machine Learning for Science

- 見過ごされていた事実の発見

- 現在は線形手法といった, 比較的単純な方法を用いていることが多く, 重要な結果が見過ごされている可能性が高い.
- 非線形手法を用いることで, 線形方法では見つけれなかったことを発見できる!
- グラフデータなどの構造化されているデータからの科学的発見

医療、材料、農業分野の研究者とタッグを組み, 提案する枠組みを既存データに適用し, 新規の科学的発見を目指す.

Hilbert-Schmidt Independence Criterion 9

Inference (hsicInf)

Yamada et al., NIPS 2016 Workshop

- 仮説検定, HSICは独立性基準
 - $H_{0,m}$: $\text{HSIC}(X_m, Y) = 0 \mid \mathcal{S} \text{ was selected,}$
 - $H_{1,m}$: $\text{HSIC}(X_m, Y) \neq 0 \mid \mathcal{S} \text{ was selected.}$
- ノンパラメトリック (非線形データを扱える)
- 多次元出力、カテゴリカルデータ等、どのようなデータにも適用可能
- p値が解析的に求まる

Post Selection Inference with Kernels

NIPS workshop 2016

& submitted to AISTATS 2018

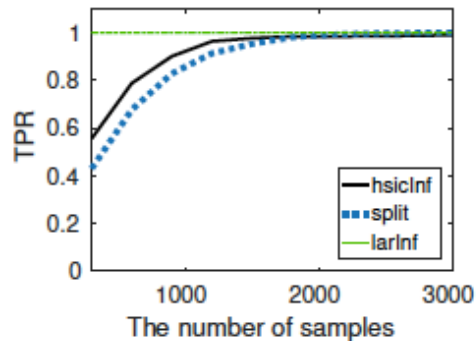
- 非線形関係のあるデータからの特徴選択 + p 値の計算

(1) 線形データ

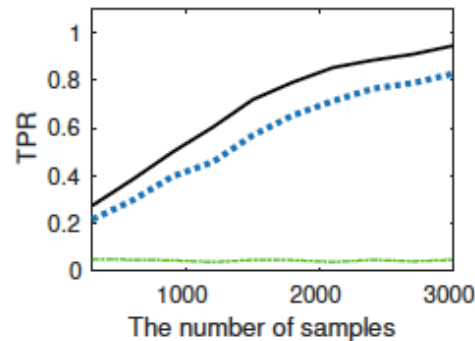
(2) 非線形データ1

(3) 非線形データ2

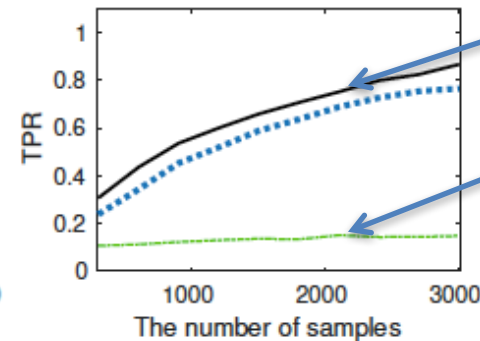
真陽性率



(a) Linear.



(b) Additive Non-linear.

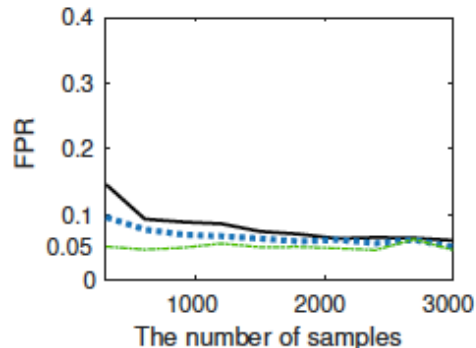


(c) Non-additive Non-linear.

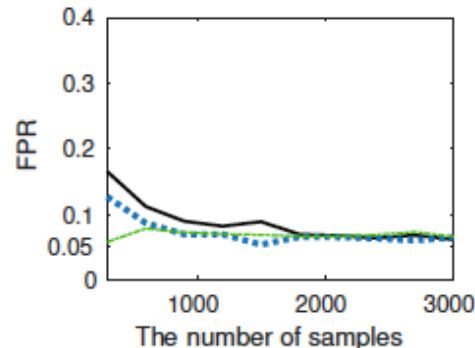
提案法: hsicInf

従来法: larInf

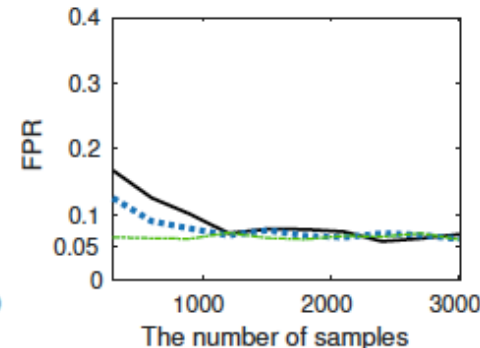
偽陽性率



(d) Linear.



(e) Additive Non-linear.



(f) Non-additive Non-linear.

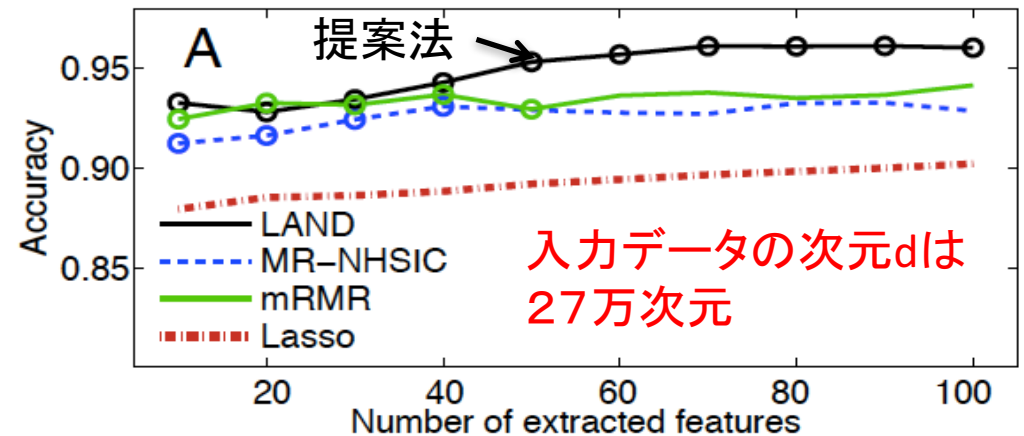
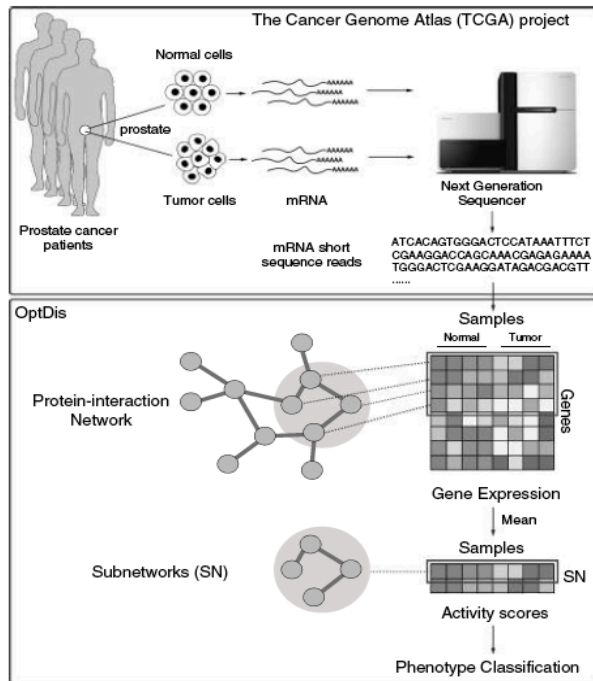
提案手法を使うことで
線形手法では見つけ
られなかった特徴を
見つけられる！

➡ 科学の研究
において重要なツール
となることが期待される

疾病予測

Disease Prediction

- 低価格での疾病予測
 - 少ない特徴数(コスト)で高い検出率



本手法を臨床の研究者が使えるように改良し,病気を高精度 & 低価格で検出できるようにする!

Hilbert-Schmidt Independence Criterion

(HSIC) Lasso (Yamada et al. NECO 2014)

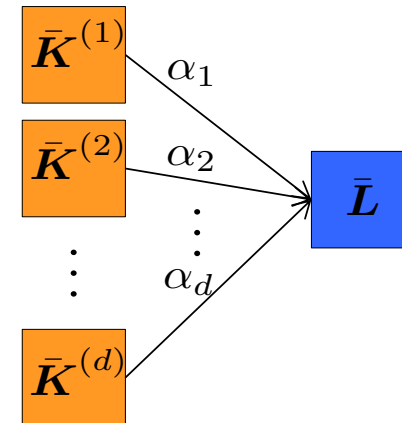
- Hilbert-Schmidt Independence Criterion Lasso

$$\min_{\alpha \in \mathbb{R}^d} \frac{1}{2} \left\| \tilde{\mathbf{L}} - \sum_{k=1}^d \alpha_k \tilde{\mathbf{K}}^{(k)} \right\|_F^2 + \lambda \|\alpha\|_1, \quad \text{s.t.} \quad \alpha_1, \dots, \alpha_d \geq 0.$$

$\tilde{\mathbf{K}}^{(k)}$: k番目の特徴のグラム行列

$\tilde{\mathbf{L}}$: 出力のグラム行列

#グラム行列 = 相関行列

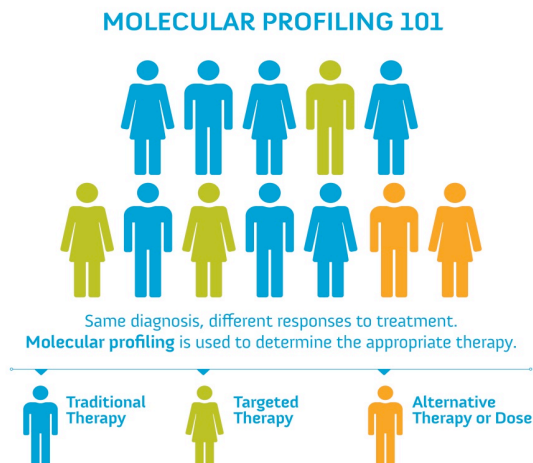


- 大域的最適解が求まる!
- 構造化データ (グラフデータ) からの特徴選択が可能
- 分散処理により, 数百万特徴 x 数万サンプル, 数万特徴, 数百万特徴から非線形に特徴を選択可能!

個別化医療

Personalized/Precision Medicine

- 個人(特定のグループ)に合わせた医療や創薬
 - 個人に効く薬の成分(特徴)を自動で見つける方法論の確立を目指す
 - 特徴選択+クラスタリングの問題として定式化



<http://sarahcannon.com/util/images/physicians/Molecular%20Profiling%20101.jpg>

THE PRECISION MEDICINE INITIATIVE



Yamada, Takeuchi, Iwata, Shawe-Taylor,
Kaski, AISTATS 2017

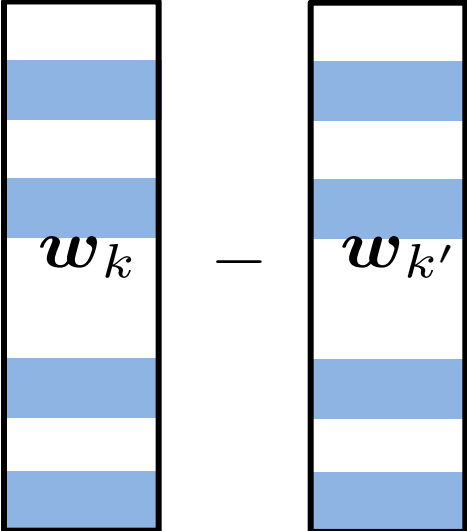
Localized Lasso (Yamada et al. AISTATS 2017)

- **ゴール:** クラスタリング+特徴選択
- 最適化問題

$$\min_{\mathbf{W}} \sum_{i=1}^n (y_i - \mathbf{w}_i^\top \mathbf{x}_i)^2 + \lambda_1 \sum_{i,j=1}^n r_{ij} \|\mathbf{w}_i - \mathbf{w}_j\|_2 + \lambda_2 \sum_{i=1}^n \|\mathbf{w}_i\|_1^2$$

クラスタリング
モデル内スパース

- アプリケーション
 - 局所回帰
 - 凸クラスタリング
 - 変化点検出



$$\mathbf{w}_k - \mathbf{w}_{k'} = \mathbf{0}$$

毒性予測

Toxicogenomics Prediction

- どのような化合物がどのような遺伝子に影響して副作用を起こすかを調べる

表1: 予測誤差 (RMSE)

	Blood			Breast			Prostate			Average
	GI50	TGI	LC50	GI50	TGI	LC50	GI50	TGI	LC50	
Localized Lasso	1.030	0.622	0.529	1.129	0.627	0.562	1.297	0.518	0.539	0.760
Network Lasso	1.096	0.918	0.921	1.368	0.821	1.065	1.475	0.711	0.690	1.007
FORMULA	1.503	1.179	1.253	1.367	1.109	1.197	1.376	1.121	1.129	1.248
Lasso	1.201	1.006	0.514	1.435	0.879	0.560	1.455	0.763	0.523	0.926
Elastic Net	1.129	0.875	0.514	1.164	0.800	0.560	1.130	0.633	0.505	0.812
Kernel Regression	1.070	0.808	0.623	1.165	0.677	0.688	1.466	0.551	0.509	0.839

低RMSE!

表2: 選択された特徴数

	Blood			Breast			Prostate			Average
	GI50	TGI	LC50	GI50	TGI	LC50	GI50	TGI	LC50	
Localized Lasso	32.7	33.4	92.5	92.6	125.9	58.2	35.1	53.9	43.4	63.4
Network Lasso	1039.6	1047.3	1052.2	1054.6	1051.5	1053.3	1060.5	1052.9	1053.1	1061.0
FORMULA	576.6	445.6	550.5	914.3	936.7	776.3	942.0	712.2	633.6	720.8
Lasso	29.6	12.0	1.0	12.0	1.9	1.0	12.5	4.4	3.8	8.7
Elastic Net	310.8	91.4	39.2	124.9	77.2	1.0	116.6	87.9	98.9	105.3

少特徴!

まとめ

Take home message

- 研究目標
 - バイオ, 医療, 材料の科学的発見をサポートする機械学習基盤の構築
 - 非線形解析を用いて, 医療, 材料, 農業分野の問題において新規科学的発見の効率化および予測性能向上を狙う
- 高次元非線形特徴選択
 - HSIC Lasso
 - 100万次元以上の特徴から非線形に特徴選択
 - hsicInf
 - 特徴選択をした上で特徴の検定をする (SI, PSI)
 - 非線形関係のあるデータの検定が可能
 - Localized Lasso
 - 個別化医療や精密医療のための数理モデリング
 - 局所的に重要な特徴を選択
- 今後の予定
 - 実応用 (バイオ、材料、etc.)
 - Beyond 非線形 (非線形性だけでは解釈は難しい?)
 - アルゴリズムのさらなる効率化