記述子設計手法
機械学習による解釈可能モデリング

兵庫県立大学高度産業科学技術研究所
藤井　将

目次

- マテリアルズ・インフォマティクス（MI）
- MIの目的
- 物質科学と機械学習の関係
- 解釈可能モデリング
- 線型独立記述子生成（LIDG）法
- LIDG法の応用例

マテリアルズ・インフォマティクス（MI）
（物質科学 × AI）

統計学、情報科学、パターン認識、データサイエンス等の分野で発達してきた機械学習（統計的学習）の手法を物質科学へ応用した研究

様々な機械学習手法

統計学

線形回帰分析
最小二乗法（OLS）
主成分分析（PCA）

統計的仮説検定
赤池情報量基準（AIC）

圧縮センシング
スパースモデリング

L1正則化（LASSO）
L2正則化（Ridge）

計算機科学

モンテカルロ法（MC）
遺伝的アルゴリズム（GA）
遺伝的プログラミング（GP）
進化的アルゴリズム（EA）
粒子群最適化（PSO）

ベイズ推定

ベイズ統計

情報科学

決定木（DT）
モンテカルロ木探索（MCTS）

ベイズ最適化

パターン認識

サポートベクターマシン（SVM
ニューラルネット（NN）
ディープラーニング（DL）
ランダムフォレスト（RF）
アンサンブル学習

MIの目的

1. 物質探索:
   良い物性値を持つ物質・材料の発見
   予測モデルの構築は必ずしも必要ではない

2. 物性予測:
   予測モデルを仮定し、未知物質の物性を高精度に予測
   非線形モデル（Black box関数）、あるいは線形モデル

3. 法則発見:
   物性発現機構の解明
   得られた知見から新たな原理、法則の発見

4. 物性データベース構築

物質科学と機械学習との関係

物質科学		機械学習
物質探索	← EA、BO、VS	最適化問題
物性予測（Black box）	← DL、RF	予測問題（Black box）
物性予測（線形モデル）	← LASSO	スパースモデリング
機構解明、法則発見	← ？？？	なし

使える統計的手法はない

→ 物質科学者が作るべき

→ 線型独立記述子生成（LIDG）法による解釈可能モデリング

解釈可能性は必要か？

社会的要請:
物質探索 ＞ 物性予測（Black box） ＞ 物性予測（解釈可能モデル）

物質探索には必ずしも物性予測は必要ではない（EA）
物性予測には必ずしも解釈可能性は必要ではない（DL、RF）

また、データ科学者の多くは効率の良い最適化や、
高速・高精度予測にしか興味がない。

解釈可能性と予測精度はある程度トレードオフの関係

80%しか当たらない解釈・説明可能なモデルと、
99%当たるブラックボックス関数では、どちらが価値が高いか？

解釈可能性の意義

・機構解明、物理法則発見につながる可能性がある
（学術的に新たな概念の発見）

・入力（記述子）と出力（物性値）との関係がわかる

ブラック・ボックス関数では入力と出力との間の関係が不明
→ どのような入力が良い出力を持つか（予め）わからない
→ また、その関数の適用範囲がわからない

一方、解釈可能なモデルであれば、どのパラメータをどれだけ上げれば物性値がどれだけ上がるか、ということがわかる。
→ 良い物質候補の見当がつく
→ モデルの適用範囲がわかる（可能性がある）
→ 限界値がわかる
→ その限界値を超えるためのヒントが得られる（可能性がある）

解釈可能なモデルとは？

1. 単純な形である
2. 経験や直感に訴える形
3. 予測精度がそこそこ高い

$$\frac{a^3}{T^2} = const.$$

よって、
解釈可能モデル（すなわち、物理的意味を抽出可能なモデル）を得るためには、モデルの構築段階で「予測精度」だけでなく、

1. 単純さ
2. 整合性

にも十分配慮する必要がある。

モデルの「単純さ」と「整合性」の定義

「単純さ」

- 説明変数が少数（スパース）である
- 説明変数の形が単純

価値が高いと判断

$$x_i, \quad \frac{x_i}{x_j}, \qquad \frac{|x_i - x_j|}{\sqrt{x_i + x_j}}$$

「整合性」

- 物理的整合性（物理的次元、対称性）
- 事前知識、経験との整合性

これらは通常の機械学習では無視されるが、LIDG法では考慮可能

線形モデルと非線形モデル

線形モデル:
- 計算コスト低い
- 過適合しにくい（low variance）
  （サンプル空間の変化に鈍感）
- シンプルなモデルが得られる
  → 物理的意味を抽出できる

- 表現力が低い（high bias）

非線形モデル:
- 計算コスト高い
- 過適合しやすい（high variance）
  （サンプル空間の変化に敏感）
- 殆どブラックボックスなモデルが得られる
  → 物理的意味の抽出は困難

- 表現力が高い（low bias）

表現力とシンプルさとの間のトレードオフをどのように解決するか？
→　記述子を新たに生成し、その中から良い記述子のみを選ぶ

解釈可能モデリングの基本方針

線形回帰　＋　記述子生成　＋　モデル選択

線形回帰分析

1. 線形モデルの仮定

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

$\vec{\varepsilon}$: 誤差項 (ガウス分布)

$$\hat{\vec{y}} = X\hat{\vec{\beta}}$$

$\hat{\vec{y}}$: 推定値

$\hat{\vec{\beta}}$: 回帰係数

目的変数
（従属変数）

他の物性値
(特徴量/説明変数/
記述子)

	$\vec{y}$	$\vec{x}_1$	$\vec{x}_2$	…	$\vec{x}_n$
サンプル 1	y_1	x_{11}	x_{12}	…	x_{1n}
サンプル 2	y_2	x_{21}	x_{22}	…	x_{2n}
…	…	…	…	…	…
サンプル m	y_m	x_{m1}	x_{m2}	…	x_{mn}

$X\ (m \times n)$:
記述子行列（計画行列）

2. ある条件のもとで回帰係数 $\hat{\vec{\beta}}$ を求める

・最小二乗法（OLS）

$$\hat{\vec{\beta}}_{OLS} = \underset{\vec{\beta}}{\operatorname{argmin}} \left\{ \left\| \vec{y} - X\vec{\beta} \right\|_{l_2}^2 \right\}$$

L_p ノルムの定義

$$\left\| \vec{\beta} \right\|_{l_p} \equiv \left(\sum_{i=1}^{n} |\beta_i|^p \right)^{1/p}$$

・LASSO

$$\hat{\vec{\beta}}_{LASSO} = \underset{\vec{\beta}}{\operatorname{argmin}} \left\{ \left\| \vec{y} - X\vec{\beta} \right\|_{l_2}^2 + \lambda \left\| \vec{\beta} \right\|_{l_1}^1 \right\}$$

記述子行列と記述子ベクトルの定義

計画行列
記述子行列
$(m \times n + 1)$

$$
\begin{array}{cc}
 & \begin{array}{ccccccc} & & & & \vec{x}_{,j} & & \end{array} \\
X = & \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\
1 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
1 & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
1 & x_{m1} & x_{m2} & \cdots & x_{mj} & \cdots & x_{mn}
\end{bmatrix} \begin{array}{c} \\ \\ \\ \vec{x}_{i,} \\ \\ \\ \end{array}
\end{array}
$$

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

パターン認識での記述子（特徴）ベクトルの定義

$$\boxed{\vec{x}_{i,} = [1, x_{i1}, x_{i2}, \dots, x_{in}]^T}$$

$n + 1$ 次元特徴空間上のベクトル
i 番目のサンプルに対する特徴量の配列

$$X = [\vec{x}_{1,}, \vec{x}_{2,}, \dots, \vec{x}_{m,}]^T$$

$$y_i = \vec{x}_{i,}{}^T \vec{\beta} + \varepsilon_i \qquad \vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

$$(y_i = \vec{\beta}^T \vec{x}_{i,} + \varepsilon_i)$$

n は通常固定
m は増加する可能性がある

スパースモデリングでの記述子べくとるの定義

$$\boxed{\vec{x}_{,j} = [x_{1j}, x_{2j}, \dots, x_{mj}]^T}$$

m 次元サンプル空間上のベクトル
j 番目の記述子の各サンプルでの値の配列

$$X = [\vec{1}, \vec{x}_{,1}, \vec{x}_{,2}, \dots, \vec{x}_{,n}]$$

$$\vec{y} = \sum_{j=0}^{n} \vec{x}_{,j} \beta_j + \vec{\varepsilon} \qquad \vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

n は増加する可能性がある
m は通常固定

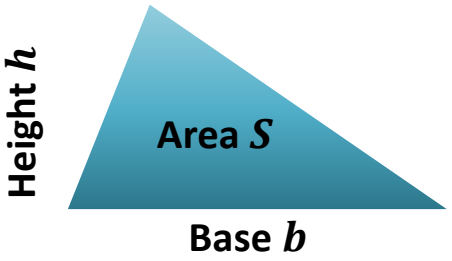
記述子生成の例

20個の様々な三角形のデータ

目的変数　　記述子

Area	Base	Height
0.0001	0.2658	0.0007
0.0092	0.9789	0.0187
0.0096	0.0488	0.3931
0.0102	0.3953	0.0514
0.0195	0.0883	0.4416
0.0273	0.2153	0.2536
0.0389	0.3864	0.2012
0.0441	0.1095	0.8048
0.0562	0.8760	0.1283
0.0577	0.2917	0.3958
0.0602	0.4255	0.2832
0.0626	0.7260	0.1725
0.0788	0.1785	0.8830
0.0871	0.7943	0.2192
0.0915	0.3709	0.4936
0.0964	0.6477	0.2976
0.1051	0.5049	0.4163
0.1119	0.7204	0.3108
0.1198	0.7600	0.3154
0.1341	0.5543	0.4839

三角形の面積



底辺と高さを記述子とした線形回帰

$$S = 0.11681\,b + 0.13780\,h - 0.03876$$

$$Adj.R^2 = 0.5515$$

底辺と高さの積を新しい記述子とした線形回帰

$$S = 0.4998\,bh - 2.333 \times 10^{-5}$$

$$Adj.R^2 = 1.0000$$

記述子の数は一つ少ないのに推定精度は良くなっている。
記述子生成の意義は、記述子の数を増加させることではなく、
良い記述子を作ることである。

このモデルから三角形の面積の法則（公式）を再発見

$$S = \frac{1}{2}bh$$

LIDGでの記述子生成1: 基本演算

基本（初期）記述子:

$$\vec{x}_1, \vec{x}_2$$

基本演算:

$$|\vec{x}_1|, |\vec{x}_2|,$$

$$\vec{x}_1^{-1}, \vec{x}_2^{-1},$$

$$|\vec{x}_1|^{-1}, |\vec{x}_2|^{-1},$$

$$|\vec{x}_1 + \vec{x}_2|, |\vec{x}_1 - \vec{x}_2|,$$

$$(\vec{x}_1 + \vec{x}_2)^{-1}, (\vec{x}_1 - \vec{x}_2)^{-1},$$

$$|\vec{x}_1 + \vec{x}_2|^{-1}, |\vec{x}_1 - \vec{x}_2|^{-1},$$

$$|\vec{x}_i| \equiv (|x_{1i}|, |x_{2i}|, \dots, |x_{mi}|)^T$$

$$\vec{x}_i^{-1} \equiv \left(x_{1i}^{-1}, x_{2i}^{-1}, \dots, x_{mi}^{-1}\right)^T$$

$$|\vec{x}_i|^{-1} \equiv \left(|x_{1i}|^{-1}, |x_{2i}|^{-1}, \dots, |x_{mi}|^{-1}\right)^T$$

$$\left|\vec{x}_i + \vec{x}_j\right| \equiv \left(\left|x_{1i} + x_{1j}\right|, \dots, \left|x_{mi} + x_{mj}\right|\right)^T$$

もちろん、もっと複雑な基本演算（平方根や指数、対数関数など）を考えても良いが、解釈性の観点から、この時点であまりにも複雑な形を考えるのは好ましくない。これだけでも、ある程度の表現力向上が見込める。

LIDGでの記述子生成2: 直積による生成

（$\vec{z}_i$は基本演算で得られた1次記述子とする。$\vec{x}_i, \vec{x}_j, 1/\vec{x}_i, |\vec{x}_i|$,...）

1st order: $\vec{z}_0,\ \vec{z}_1,\ \vec{z}_2$ $\quad ({}_3H_1 = 3)$

2nd order: $\vec{z}_0^2,\ \vec{z}_1^2,\ \vec{z}_2^2,$
$\vec{z}_0\vec{z}_1,\ \vec{z}_0\vec{z}_2,\ \vec{z}_1\vec{z}_2$ $\quad ({}_3H_2 = 6)$

3rd order: $\vec{z}_0^3,\ \vec{z}_1^3,\ \vec{z}_2^3,$
$\vec{z}_0^2\vec{z}_1,\ \vec{z}_0^2\vec{z}_2,\ \vec{z}_1^2\vec{z}_2,$
$\vec{z}_0\vec{z}_1^2,\ \vec{z}_0\vec{z}_2^2,\ \vec{z}_1\vec{z}_2^2,$
$\vec{z}_0\vec{z}_1\vec{z}_2$ $\quad ({}_3H_3 = 10)$

4th order: ... $\quad ({}_3H_4 = 15)$

5th order: ... $\quad ({}_3H_5 = 21)$

$$\vec{z}_i\vec{z}_j \equiv \left(z_{1i}z_{1j}, z_{2i}z_{2j}, \dots, z_{mi}z_{mj}\right)^T$$

このようにして、記述子を必要なだけ段階的にシステマティックに生成できる。
記述子の複雑さも段階的に増加させることができる。
（もちろんこれらの記述子の間には後述する多重共線性が生じる可能性がある）

多重共線性

共線性 $\vec{x}_i = c_j\vec{x}_j + c_0$

準共線性 $\vec{x}_i \sim c_j\vec{x}_j + c_0$

多重共線性 $\vec{x}_i = c_j\vec{x}_j + c_k\vec{x}_k + \cdots + c_0$

準多重共線性 $\vec{x}_i \sim c_j\vec{x}_j + c_k\vec{x}_k + \cdots + c_0$

記述子生成を行うと、容易にこのような関係式が生成されてしまう。

何故、多重共線性を検出・除去しなければならないのか？
（何故、記述子空間を線形独立化しなければならないか？）

- 回帰係数の増大、発散
- サンプル空間の変化に対するモデルの不安定性（汎化性能の低下）
- モデル選択の不安定性

最小二乗法 (OLS)

$$X = [\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_n]$$

$$\hat{\vec{\beta}}_{\mathrm{OLS}} = \underset{\vec{\beta}}{\mathrm{argmin}} \left\{ \left\| \vec{y} - X\vec{\beta} \right\|_{l_2}^2 \right\} \qquad L(\vec{\beta}) \equiv \left\| \vec{y} - X\vec{\beta} \right\|_{l_2}^2$$

$$\frac{\partial L(\vec{\beta})}{\partial \vec{\beta}} = -\left[X^T\vec{y} - X^T X\vec{\beta}\right] = \vec{0}$$

$$\hat{\vec{\beta}}_{\mathrm{OLS}} = \left(X^T X\right)^{-1} X^T \vec{y} \qquad \left(X^T X\right)^{-1} = \frac{\Delta}{\det|X^T X|}$$

但し、$X^T X$ は正則行列であるとする。
つまり、 $X^T X$ は逆行列を持つとする。

$$\det\left|X^T X\right| \neq 0$$

→ これはつまり、$\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_n$ が線型独立であるということ。

正則化の3つのアプローチ

$X = [\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_n]$ において、
$\vec{x}_1, \vec{x}_2, \vec{x}_3, \dots, \vec{x}_n$ が線形独立でなければ？

1. 罰則項を導入し正則化して、緩和問題を解く(Ridge, LASSO)
2. 基底変換により、直交基底を作る (PCR, PLS)
3. 多重共線性関係を全て見つけて除去し、線型独立化させる(LIDG)

多重共線性

$$\vec{x}_i = c_1\vec{x}_1 + c_2\vec{x}_2 + \cdots + c_n\vec{x}_n$$

$$\vec{0} = X\vec{c}$$

つまり、多重共線性が存在するということは、$X\vec{c} = \vec{0}$ が非自明な解を持つということ。
→ よって、多重共線性を検出するためには、その非自明解（のセット）を見つければ良い。

$X\vec{c} = \vec{0}$ が非自明な解を持つ条件:

1. 劣決定系、即ち $m < n$ のとき
2. ランクが落ちているとき、即ち $\mathrm{rank}(X) \equiv r < \min(m, n)$

$$\begin{array}{cc} & n \\ m & \begin{bmatrix} 1 & 0 & 0 & 0 & a \\ 0 & 1 & 0 & 0 & b \\ 0 & 0 & 1 & 0 & c \\ 0 & 0 & 0 & 1 & d \end{bmatrix} \end{array} \qquad \begin{array}{cc} & n \\ m & \begin{bmatrix} 1 & 0 & 0 & a \\ 0 & 1 & 0 & b \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

非自明解の数は以下で与えられる

$$n - r$$

これが記述子間の多重共線性関係の数

部分空間リスト選択法による多重共線性の除去

階段行列法から以下の多重共線性が得られたとする

$$\vec{x}_k = c_i^k \vec{x}_i + c_j^k \vec{x}_j$$
$$\vec{x}_l = c_i^l \vec{x}_i + c_j^l \vec{x}_j$$
$$\vec{x}_m = c_i^m \vec{x}_i + c_j^m \vec{x}_j$$

→

リスト表示

$$\left[\vec{x}_k, \vec{x}_l, \vec{x}_m, \left[\vec{x}_i, \vec{x}_j\right]\right]$$

Temporary basis

この線形包 $\mathrm{span}(\{\vec{x}_i, \vec{x}_j, \vec{x}_k, \vec{x}_l, \vec{x}_m\})$
の基底の数（次元）は 2 しかない

$$W = \mathrm{span}(\{\vec{x}_i, \vec{x}_j, \vec{x}_k, \vec{x}_l, \vec{x}_m\})$$

$$X' = [\vec{x}_i, \vec{x}_j, \vec{x}_k, \vec{x}_l, \vec{x}_m]$$

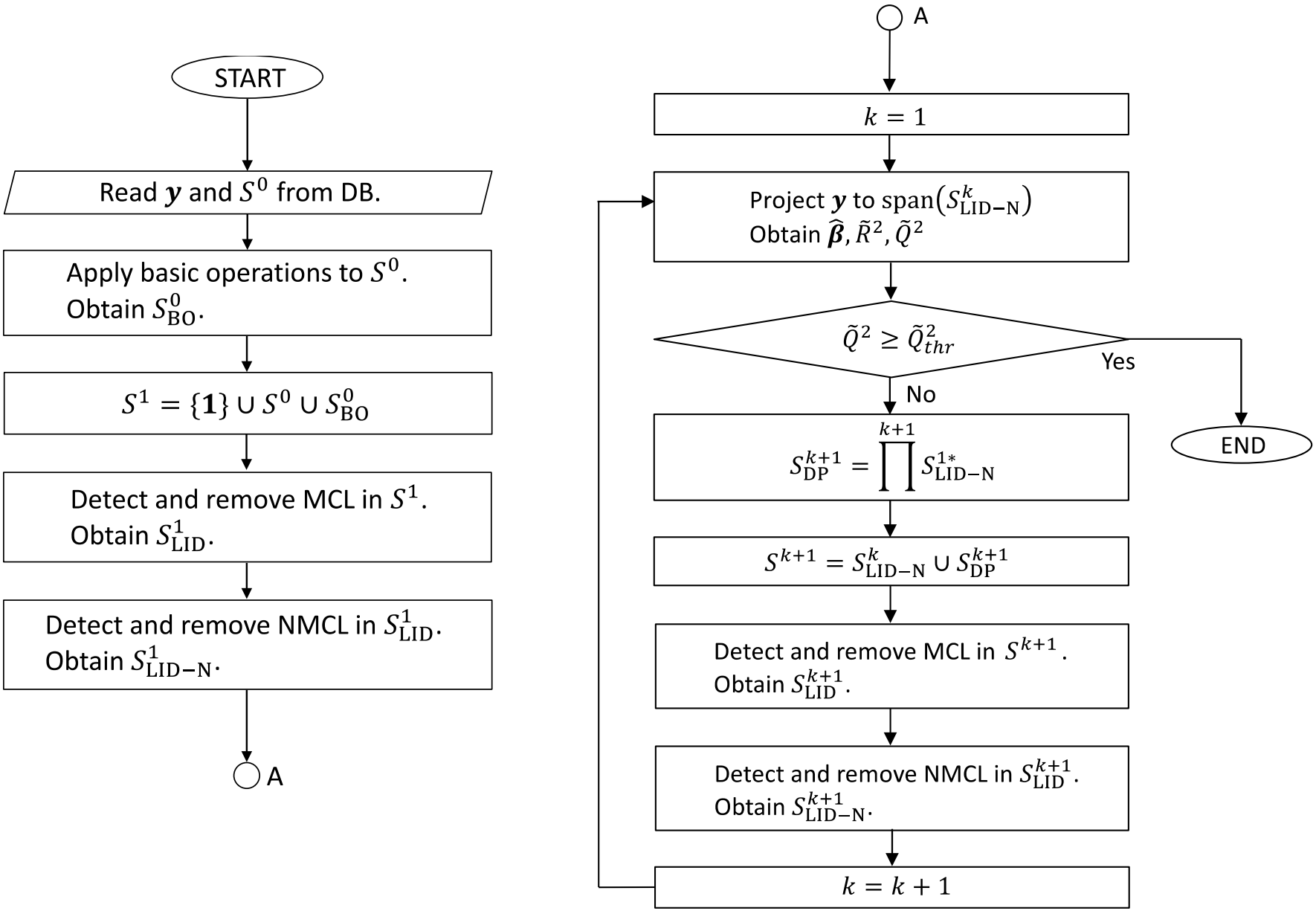
$$\dim(W) = \mathrm{rank}(X') = 2$$

よって、上のリストから好きな記述子を 2 つ選ぶようにする。

select $\vec{x}_i, \vec{x}_j$ and remove $\vec{x}_k, \vec{x}_l, \vec{x}_m$
or
select $\vec{x}_i, \vec{x}_k$ and remove $\vec{x}_j, \vec{x}_l, \vec{x}_m$
or
select $\vec{x}_l, \vec{x}_m$ and remove $\vec{x}_i, \vec{x}_j, \vec{x}_k$
…

初期記述子空間が持つ
いかなる情報（回帰精度）も落とさずに記述子の数を減らすことができる。

LIDG method


START
Read $\boldsymbol{y}$ and S^0 from DB.
Apply basic operations to S^0.
Obtain S^0_{BO}.
$S^1 = \{\mathbf{1}\} \cup S^0 \cup S^0_{\mathrm{BO}}$
Detect and remove MCL in S^1.
Obtain S^1_{LID}.
Detect and remove NMCL in S^1_{LID}.
Obtain $S^1_{\mathrm{LID-N}}$.
A
A
$k = 1$
Project $\boldsymbol{y}$ to $\mathrm{span}(S^k_{\mathrm{LID-N}})$
Obtain $\widehat{\boldsymbol{\beta}}, \tilde{R}^2, \tilde{Q}^2$
$\tilde{Q}^2 \geq \tilde{Q}^2_{thr}$
Yes
No
END
$S^{k+1}_{\mathrm{DP}} = \prod^{k+1} S^{1*}_{\mathrm{LID-N}}$
$S^{k+1} = S^k_{\mathrm{LID-N}} \cup S^{k+1}_{\mathrm{DP}}$
Detect and remove MCL in S^{k+1}.
Obtain S^{k+1}_{LID}.
Detect and remove NMCL in S^{k+1}_{LID}.
Obtain $S^{k+1}_{\mathrm{LID-N}}$.
$k = k + 1$

先行研究の紹介
LIDGの適用例、先行研究との比較

先行研究

L. M. Ghiringhelli, *et al.,* PRL 114, 105503 (2015)

1. ある少数の初期記述子に対し、四則演算や和や差の絶対値や指数など、あらゆる演算を施して高次元記述子空間を作る

2. その後LASSO（least absolute shrinkage and selection operator）と呼ばれる罰則項付きの線形回帰によって少数の記述子に絞り込む

これにより、
表現力、汎化性能が高く、且つ、比較的シンプルな回帰モデルを抽出することに成功

$$\Delta E \equiv E_{AB}(RS) - E_{AB}(ZB)$$

$$= 0.108\frac{EA(B) - IP(B)}{r_p^2(A)} + 1.79\frac{|r_s(A) - r_p(B)|}{exp(r_s(A))} + 3.766\frac{|r_p(B) - r_s(B)|}{exp(r_d(A))} - 0.0267$$

先行研究の問題点

1. 得られた回帰式がまだ複雑
　分母に指数関数が含まれていたりして、直ぐには物理的意味を抽出することができない

2. LASSOによるモデル選択においては、記述子の間に強い相関がある場合、どの記述子が選ばれるかコントロールできない

3. （記述子生成の部分がシステマティックでない）

Previous work

Sample space setting

$m \sim 100$

Calculation of $\vec{y}$

Prep. basic descriptors

$X^{(0)} = [\vec{1}, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n]$

$n \sim 10$

Descriptor generation

X

$n \sim 1000$

LASSO

$n \sim 10$

L. M. Ghiringhelli, *et al.*, PRL 114, 105503 (2015)

Our approach

Sample space setting

$m \sim 100$

Calculation of $\vec{y}$

Prep. basic descriptors

$X^{(0)} = [\vec{1}, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n]$

$n \sim 10$

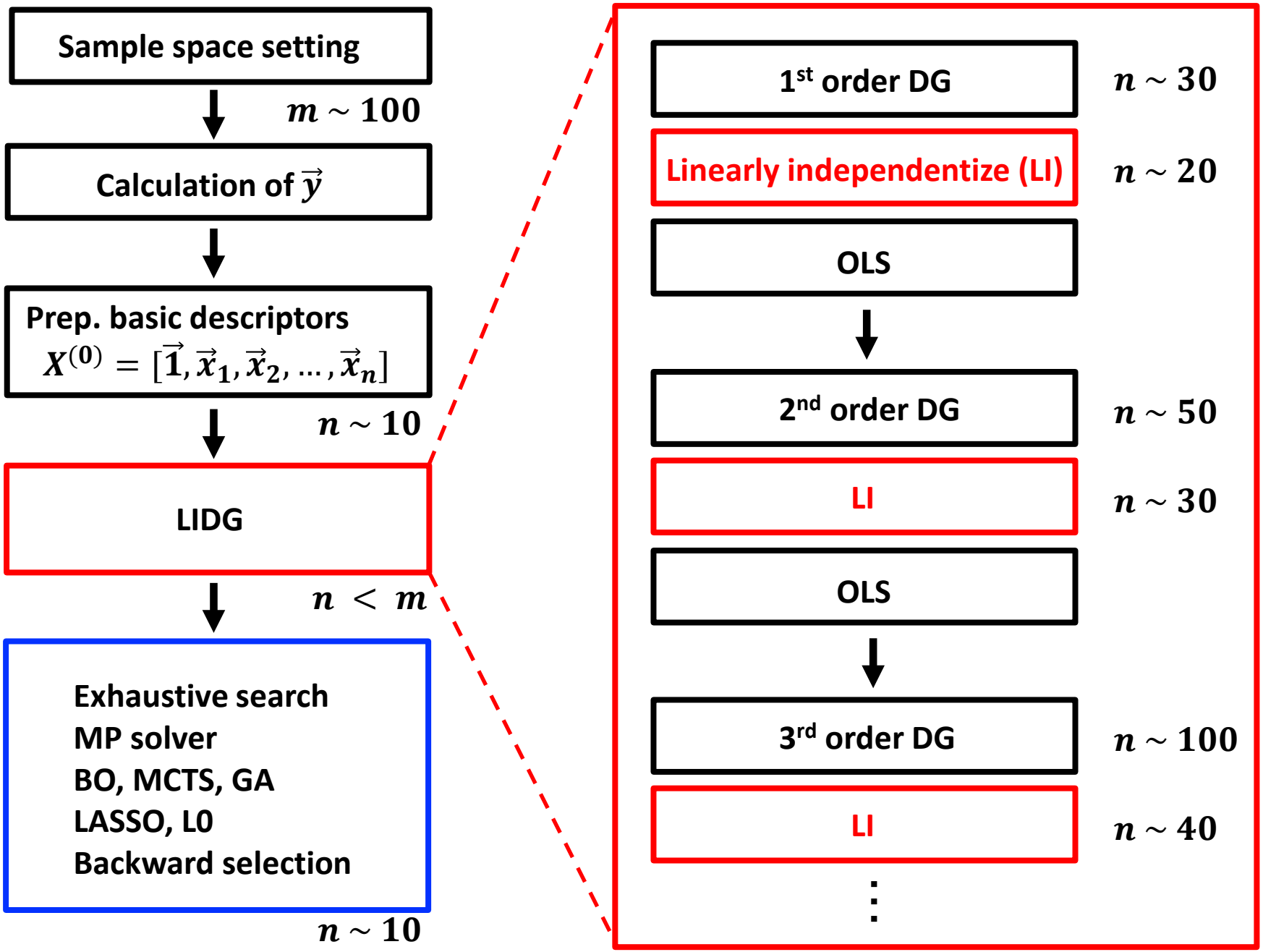
Linearly independent descriptor generation (LIDG)

$n < m$

Exhaustive search
MP solver
BO, MCTS, GA
LASSO, L0
Backward selection

$n \sim 10$

Linearly independent descriptor generation method



適用例

目的変数: 2元半導体（82個）のRS構造とZB構造のエネルギー差

$$\Delta E = E(\mathrm{RS}) - E(\mathrm{ZB}).$$

$$\Delta \boldsymbol{E}(\boldsymbol{A}, \boldsymbol{B}) = \Delta \boldsymbol{E}(\boldsymbol{B}, \boldsymbol{A})$$

対称化記述子（A元素、B元素の原子半径）

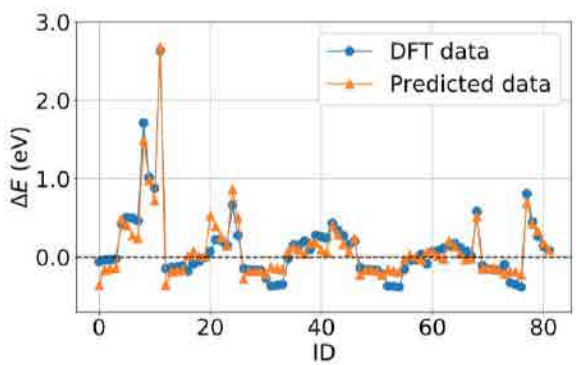
$$x_A + x_B, \quad |x_A - x_B|, \quad |x_A + x_B|, \quad \frac{1}{x_A + x_B}, \quad \frac{1}{|x_A + x_B|}.$$

回帰結果

$$\begin{aligned} \Delta E &= 6.87 \left(\frac{1}{r_A + r_B} \right)^3 - 5.02 \frac{|r_A - r_B|}{(r_A + r_B)^3} - 0.18 \\ &= \left(\frac{1}{r_A + r_B} \right)^3 \{-5.02 |r_A - r_B| + 6.87\} - 0.18 \end{aligned}$$

先行研究のモデル

$$\frac{IP_B - EA_B}{r_{pA}^2}, \quad \frac{|r_{sA} - r_{pB}|}{\exp(r_{sA})}, \quad \frac{|r_{pB} - r_{sB}|}{\exp(r_{dA})}.$$



	Present		Previous work [5]		
Criterion	Model 1	Model 2	Model A	Model B	Model C
M	3	2	1	2	3
R^2	0.913	0.876	0.883	0.929	0.957
Q^2	0.902	0.866	0.867	0.918	0.946
AIC	−92.4	−65.0	−72.0	−110.6	−149.4
MAE (eV)	0.102	0.118	0.121	0.097	0.071
MaxAE (eV)	0.457	0.460	0.400	0.349	0.301

Y. Kanda, H. Fujii, and T. Oguchi, STAM 20, 1178 (2019).

先行研究との比較、LIDG法の優位性

・記述子を段階的に増加させることができる。
また、記述子を段階的に複雑なものにすることができる。

・記述子を必要な数だけ生成することができる。（後のモデル選択が楽になる）

・記述子選択に研究者の知識・経験をある程度導入することができる。
（部分空間リスト）

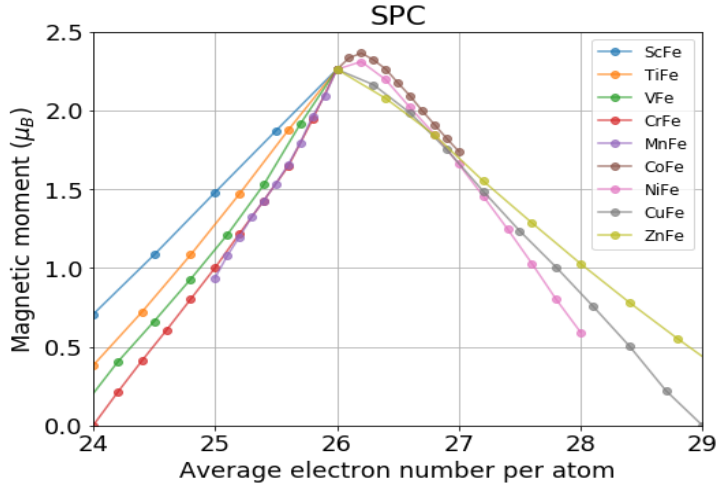
・多重共線性や強い準多重共線性が取り除かれているので、
→　回帰係数が無用に大きくならず、有意な回帰分析を行うことができる
→　安定したモデル選択が可能
（最も単純な後退選択法も、LASSOも同様のモデルを選択する）

・完全な多重共線性関係が得られるので、それを使って、
最終的に得られたモデルの式変形を行うことができる。

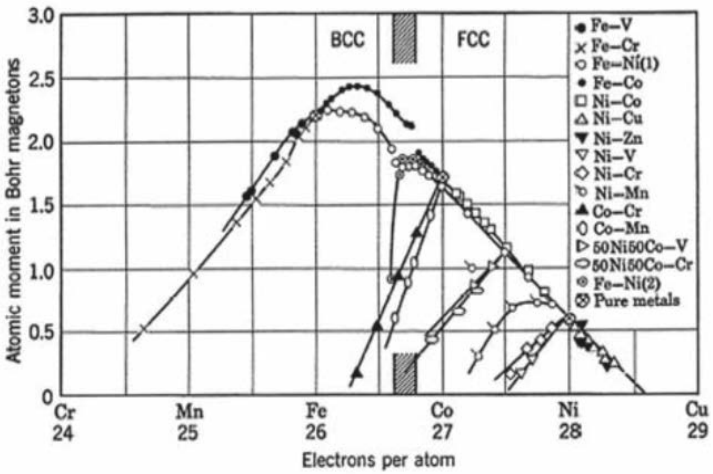
Example

AkaiKKR (KKR-CPA)
LDA (MJW)
Crystal structure: BCC
Lattice constant = 2.86 Å
Scalar relativistic
BZ quality = 10 (nk=256)
edelta = 0.0001
ewidth = 1.2
Complex energy mesh = 85
L max = 3

Calculation



Experiment



H. Saito: Physics and Applications of Invar Alloys, Maruzen, Tokyo, (1978), 18.

Everyone may image that there is a simple model on the back of this curve
We try to find the simple model by using LIDG method

Target property ($m = 99$):
Magnetic moment of binary alloy $A_{1-x}B_x$

$$M(A, B, x)$$

Initial descriptors ($n = 7$):
Concentration, x
Magnetic moment of pure bulk, $M_P(A), M_P(B)$
The number of valence electrons, $Z(A), Z(B)$
Magnetic moment of impurity atom, $M_I(A, B)$
$M_I(B, A)$

The results of OLS with 1st, 2nd, and 3rd order descriptors

1st order (8)

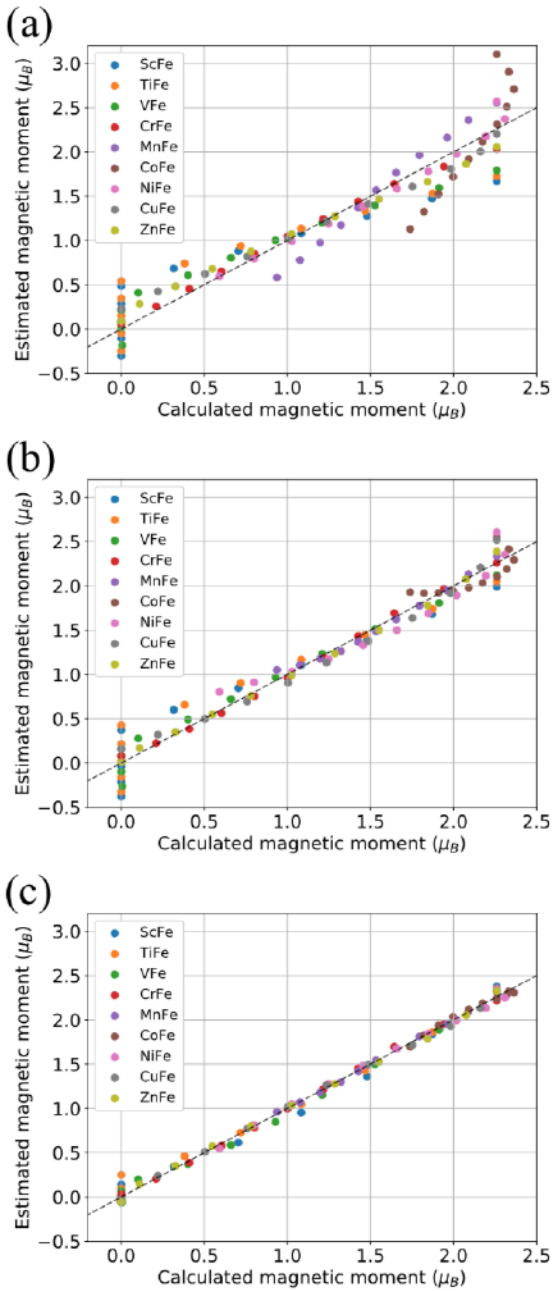
$$R^2 = 0.906$$
$$Q^2 = 0.887$$

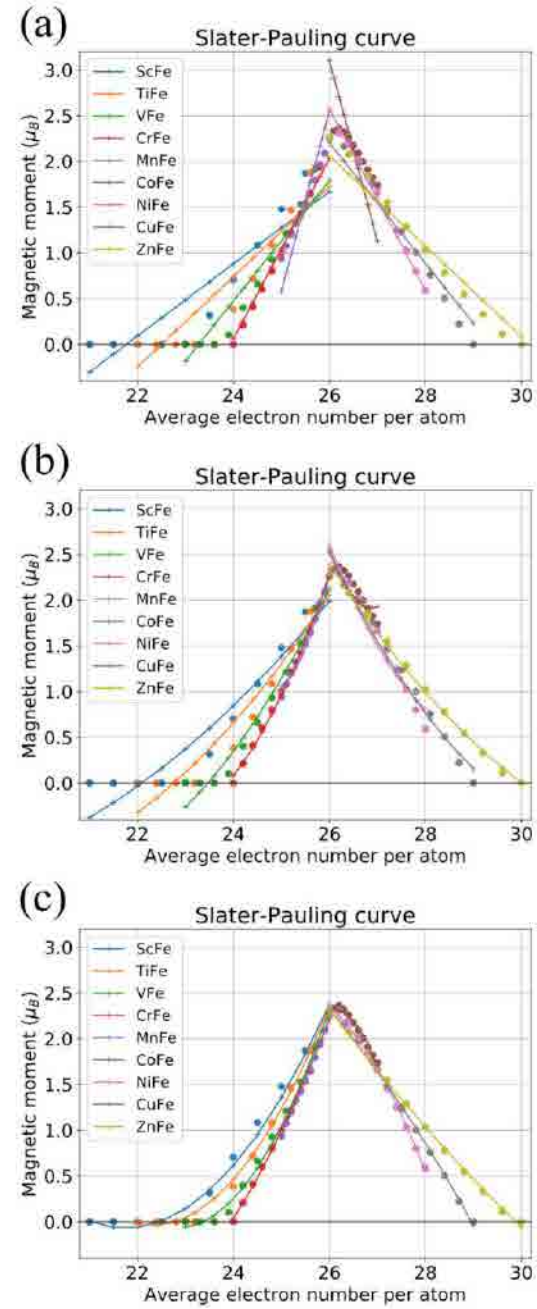
Up to 2nd order (17)

$$R^2 = 0.969$$
$$Q^2 = 0.950$$

Up to 3rd order (24)

$$R^2 = 0.996$$
$$Q^2 = 0.991$$





Obtained model by backward selection

$$\boldsymbol{M(A,B,x) \sim (1-x)M_P(A) + xM_P(B)}$$
$$\boldsymbol{+x(1-x)[-4.03 + 0.28Z(A) + 0.62M_I(B,A) + 0.20|M_P(A) + M_I(A,B)|]}$$

An interpretation
by analogy with the regular solution approximation for binary compound system

$$\boldsymbol{M(A,B,x) = \bar{M}(A,B,x) + \Delta M(A,B,x)}$$

$$\boldsymbol{\bar{M}(A,B,x) \equiv (1-x)M_P(A) + xM_P(B)}$$

Averaged magnetic moment term (linear to x)

$$\boldsymbol{\Delta M(A,B,x) \equiv x(1-x)\Omega(A,B)}$$

Mixing (excessive) magnetic moment term generated by alloying

$$\boldsymbol{\Omega(A,B) = -4.03 + 0.28Z(A) + 0.62M_I(B,A) + 0.20|M_P(A) + M_I(A,B)|}$$

Interaction parameter (does not depend on x)

Summary

- **We propose RREF method for detecting MCL**
- **The subspace list is useful for breaking the detected MCL relationships**
- **This is a new approach to solve MCL problem in linear regression analysis**
- **By combining these methods, LIDG method was proposed**
- **LIDG method was applied to analyze SPCs and we could obtain a simple (interpretable) model with high generalization capability.**

Use of symmetry in target property

If we can find symmetries,

1. **we can increase samples which are symmetrically equivalent.**
2. **we can construct descriptors to satisfy the above regulations in advance**

Magnetic moment of alloys expected to have a symmetry like,

$$M(A, B, x) = M(B, A, 1 - x)$$

$$M(A, B, x_A, x_B) = M(B, A, x_B, x_A)$$

A, B, x_A, x_B: **identifiers (primary key) (not descriptor)**

$$EM(A, B, x_A, x_B) = M(A, B, x_A, x_B)$$
$$\sigma M(A, B, x_A, x_B) = M(B, A, x_B, x_A) = M(A, B, x_A, x_B)$$

Symmetrization operator (projection operator):

$$S = E + \sigma$$

Symmetrization of descriptors

Symmetrization operator:

$$\boldsymbol{S = E + \sigma}$$

$$\boldsymbol{SM_P(A) = M_P(A) + M_P(B)}$$
$$\boldsymbol{SM_P(B) = M_P(B) + M_P(A)}$$

$$\boldsymbol{Sx_A = x_A + x_B}$$

$$\boldsymbol{Sx_B = x_B + x_A}$$

$$\boldsymbol{SM_I(A,B) = M_I(A,B) + M_I(B,A)}$$

$$\boldsymbol{SM_I(B,A) = M_I(B,A) + M_I(A,B)}$$

$$\boldsymbol{Sx_AM_P(A) = x_AM_P(A) + x_BM_P(B)}$$

$$M(A,B,x) = \overline{M}(A,B,x) + \Delta M(A,B,x),$$

$$\overline{M}(A,B,x) \equiv (1-x)M_{\mathrm{P}}(A) + xM_{\mathrm{P}}(B),$$

$$\Delta M(A,B,x) \equiv x(1-x)\Omega(A,B).$$

If we use these symmetrical descriptors, we can get a more reasonable model.

The number of multicollinearities

Linear span:

$$\boldsymbol{W} = \mathbf{span}(\{\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2, \vec{\boldsymbol{x}}_3, \dots, \vec{\boldsymbol{x}}_n\})$$

Design matrix:

$$\boldsymbol{X} = [\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2, \vec{\boldsymbol{x}}_3, \dots, \vec{\boldsymbol{x}}_n]$$

The number of non-trivial solutions of $\boldsymbol{X}\vec{\boldsymbol{c}} = \vec{\mathbf{0}}$
means the number of extra basis in span $\boldsymbol{W}$

$$\boldsymbol{n} - \mathbf{dim}(\boldsymbol{W})$$

$$\mathbf{dim}(\boldsymbol{W}) = \mathbf{rank}(\boldsymbol{X}) \equiv \boldsymbol{r}$$

then, the number of non-trivial (independent) solutions is given by $\boldsymbol{n} - \boldsymbol{r}$

independent one set of solution

RREF method

Make row reduced echelon form (rref) of X by basic operations.

$$RXQ = X_{rref}$$

$R(m \times m)$: **row basic transformation operator (regular)**

$Q(n \times n)$: **column basic transformation operator (regular and orthogonal)**
(here suppose that it just change the order of columns)

$$X_{rref} = \begin{bmatrix} 1 & 0 & 0 & a_1 & b_1 & d_1 \\ 0 & 1 & 0 & a_2 & b_2 & d_2 \\ 0 & 0 & 1 & a_3 & b_3 & d_3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$n - r = 6 - 3 = 3$$

$$X_{rref}[\vec{c}_1, \vec{c}_2, \vec{c}_3] = [\vec{0}, \vec{0}, \vec{0}]$$

The simplest solutions:

$$[\vec{c}_1, \vec{c}_2, \vec{c}_3] = \begin{bmatrix} -a_1 & -b_1 & -d_1 \\ -a_2 & -b_2 & -d_2 \\ -a_3 & -b_3 & -d_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & a_1 & b_1 & d_1 \\ 0 & 1 & 0 & a_2 & b_2 & d_2 \\ 0 & 0 & 1 & a_3 & b_3 & d_3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -a_1 \\ -a_2 \\ -a_3 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \vec{0}$$

RREF method

The solutions of $X_{rref}\vec{c} = \vec{0}$ are the solutions of $X\vec{c} = \vec{0}$?

$$RXQ = X_{rref} \qquad X = R^{-1}X_{rref}Q^T$$

$$X_{rref}\vec{c} = \vec{0}$$
$$R^{-1}X_{rref}\vec{c} = \vec{0}$$
$$R^{-1}X_{rref}Q^TQ\vec{c} = \vec{0}$$
$$X\vec{Q\vec{c}} = \vec{0}$$

How to obtain Q

We only have to remember the order change

$$XQ = [\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4]Q = [\vec{x}_3, \vec{x}_1, \vec{x}_2, \vec{x}_4]$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} Q = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$Q = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Raw data

Label	$M(A,B,x)$	x	$M_{\mathrm{P}}(A)$	$Z(A)$	$M_{\mathrm{P}}(B)$	$Z(B)$	$M_{\mathrm{I}}(A,B)$	$M_{\mathrm{I}}(B,A)$
Sc100Fe000	0.0000	0.0	0.0000	3	2.2604	8	-0.32667	0.00000
Sc090Fe010	0.0000	0.1	0.0000	3	2.2604	8	-0.32667	0.00000
Sc080Fe020	0.0000	0.2	0.0000	3	2.2604	8	-0.32667	0.00000
Sc070Fe030	0.0000	0.3	0.0000	3	2.2604	8	-0.32667	0.00000
Sc060Fe040	0.0000	0.4	0.0000	3	2.2604	8	-0.32667	0.00000
Sc050Fe050	0.3157	0.5	0.0000	3	2.2604	8	-0.32667	0.00000
Sc040Fe060	0.7056	0.6	0.0000	3	2.2604	8	-0.32667	0.00000
Sc030Fe070	1.0831	0.7	0.0000	3	2.2604	8	-0.32667	0.00000
Sc020Fe080	1.4793	0.8	0.0000	3	2.2604	8	-0.32667	0.00000
Sc010Fe090	1.8709	0.9	0.0000	3	2.2604	8	-0.32667	0.00000
Sc000Fe100	2.2604	1.0	0.0000	3	2.2604	8	-0.32667	0.00000
Ti100Fe000	0.0000	0.0	0.0000	4	2.2604	8	-0.69382	0.00000
Ti090Fe010	0.0000	0.1	0.0000	4	2.2604	8	-0.69382	0.00000
Ti080Fe020	0.0000	0.2	0.0000	4	2.2604	8	-0.69382	0.00000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Zn040Fe060	1.2862	0.6	0.0000	12	2.2604	8	0.03119	0.00000
Zn030Fe070	1.5524	0.7	0.0000	12	2.2604	8	0.03119	0.00000
Zn020Fe080	1.8440	0.8	0.0000	12	2.2604	8	0.03119	0.00000
Zn010Fe090	2.0763	0.9	0.0000	12	2.2604	8	0.03119	0.00000
Zn000Fe100	2.2604	1.0	0.0000	12	2.2604	8	0.03119	0.00000

Descriptors

M_P: Magnetic moment of pure bulk A
Z: # of valence electrons

A	$M_P(A)$	$Z(A)$
Sc	0.000	3
Ti	0.000	4
V	0.005	5
Cr	0.000	6
Mn	0.937	7
Fe	2.260	8
Co	1.740	9
Ni	0.591	10
Cu	0.000	11
Zn	0.000	12

M_I(A,B):
Magnetic moment of impurity A (B) in bulk B (A)

A	B	$M_I(A,B)$	$M_I(B,A)$
Sc	Fe	-0.327	0.000
Ti	Fe	-0.694	0.000
V	Fe	-1.165	-0.005
Cr	Fe	-1.644	1.825
Mn	Fe	-1.769	2.387
Co	Fe	1.846	2.673
Ni	Fe	1.075	2.797
Cu	Fe	0.268	2.332
Zn	Fe	0.031	0.000

放射光を利用したマテリアルズ・インフォマティクス
計測インフォマティクス

放射光インフォマティクス

物質・材料科学の目標

物質（や、その作成条件）→ 目的物性値

サンプル物質から、目的となる物性値を高速・高精度測定、予測
（所望の物性値を持つ物質の探索の高効率化）

放射光実験、DFT計算の役割

実験の（時間、費用）コストダウン
メカニズムの解明（仮説、ストーリーの検証）

（ただし、スペクトル → 目的物性値　が既知の場合）

放射光インフォマティクス

物質 → 目的物性値

実験による高速・高精度測定が困難な場合

DFT計算

物質 → 電子状態 → 目的物性値

放射光

物質 → スペクトル → 目的物性値

問題点：

1. どういう物質（の状態）を見ているか不確かな場合がある
2. スペクトル、DFT計算の結果の解釈が難しい場合がある
3. スペクトル取得、DFT計算のコストが高い場合がある

放射光インフォマティクス

機械学習でやりたいこと:　スペクトル解析器、予測器の構築

（実験・計算）
スペクトル

→

解析器
予測器

→

1. 物質の構造
2. 電子状態
3. 物性値

・スペクトル解釈性の向上（酸化数、配位数、局所電子状態）
・過去のデータから最適な実験の提案（L端？K端？XAS？XPS？）
・ユーザーの利便性の向上
・新分野開拓

RIXS-MCDの解釈、XASスペクトル等の高エネルギー領域の解釈（Co K端）
今まで無視していた部分から何らかの意味が抽出されるかもしれない。

その他

（実験・計算）スペクトルデータベース構築
データ補完、ノイズ除去による精度向上

Figure 2. Random impulse noise. Our denoiser is trained on corrupted image pairs only.

NVIDIA, arXiv:1803.04189v3 [cs.CV] 29 Oct 2018