全状態探索型スパースモデリング
ES-SpM
(Exhaustive Search for Sparse Modeling)

岡田真人

東京大学　大学院新領域創成科学研究科
複雑理工学専攻

© 2024 Masato Okada

本解説の主張

- データ駆動科学における特徴量選択や記述子抽出では，全状態探索型のスパースモデリングを用いるのが適切である.

内容

- 機械学習と特徴量選択
- 全状態探索法とWeight Diagram
- 特徴量選択の信頼度評価
- 拡張された特徴量空間における特徴量選択

変数選択に関する二つの戦略

原則：変数選択の問題の計算量は指数爆発する
(Cover and Van Campenhout, 1977)

＜変数選択に対する二つの戦略＞

1. 凸最適化や変分ベイズにもとづく緩和型アプローチ
   通常に用いられているスパース推定のアルゴリズム
   凸最適化: Lasso(1996,L1正則化)
   変分ベイズ: ARD(2008,関連自動度決定)

2. 全状態を効率的に探索(サンプリングアプローチ)
   サンプリング：MCMC(1993,マルコフ連鎖モンテカルロ法)
   REMC(2006,レプリカ交換モンテカルロ法)

データ駆動科学には全状態探索型SpM

線形回帰モデル

- 目的変数：y
- 準備した特徴量：$(x_1, x_2, \ldots, x_p)$

関係式 $y = g(x_1, x_2, \ldots, x_p)$ に対して，線形和による近似を考える

$$y = g(x_1, x_2, \ldots, x_p)$$
$$\approx w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

ただし w_i は回帰係数

予測モデルの汎化性能

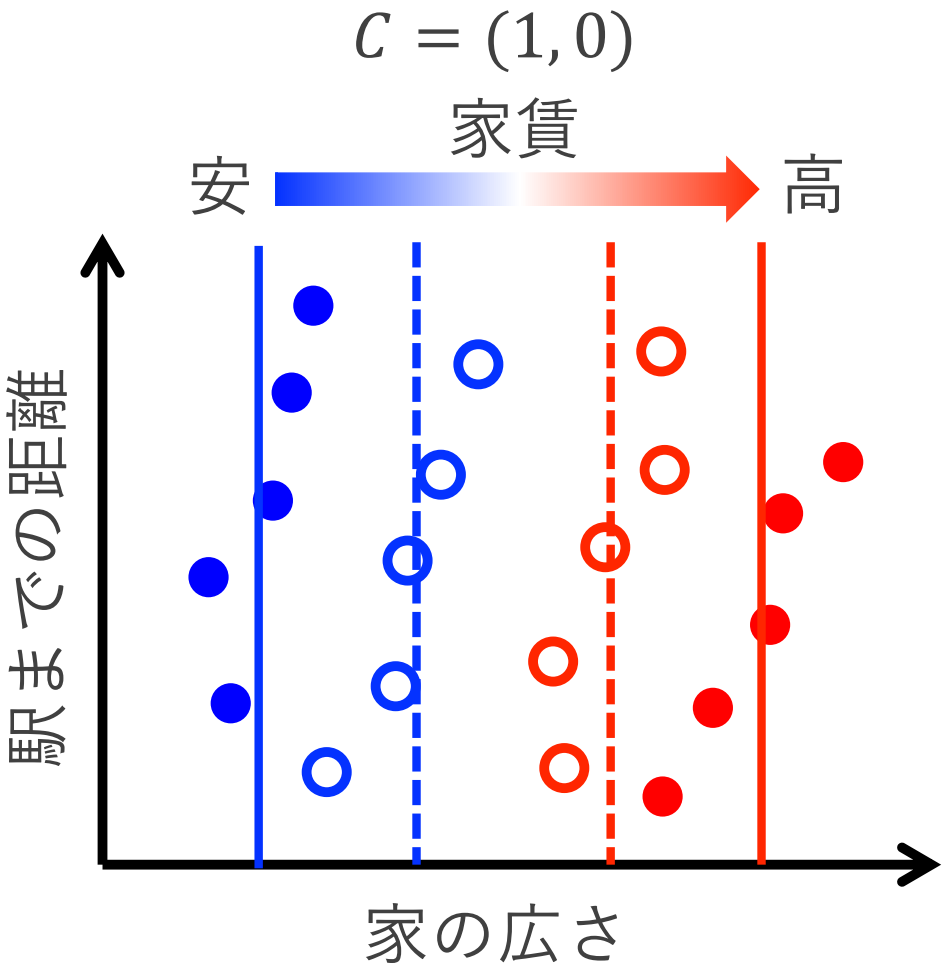
機械学習モデルに求められる性質

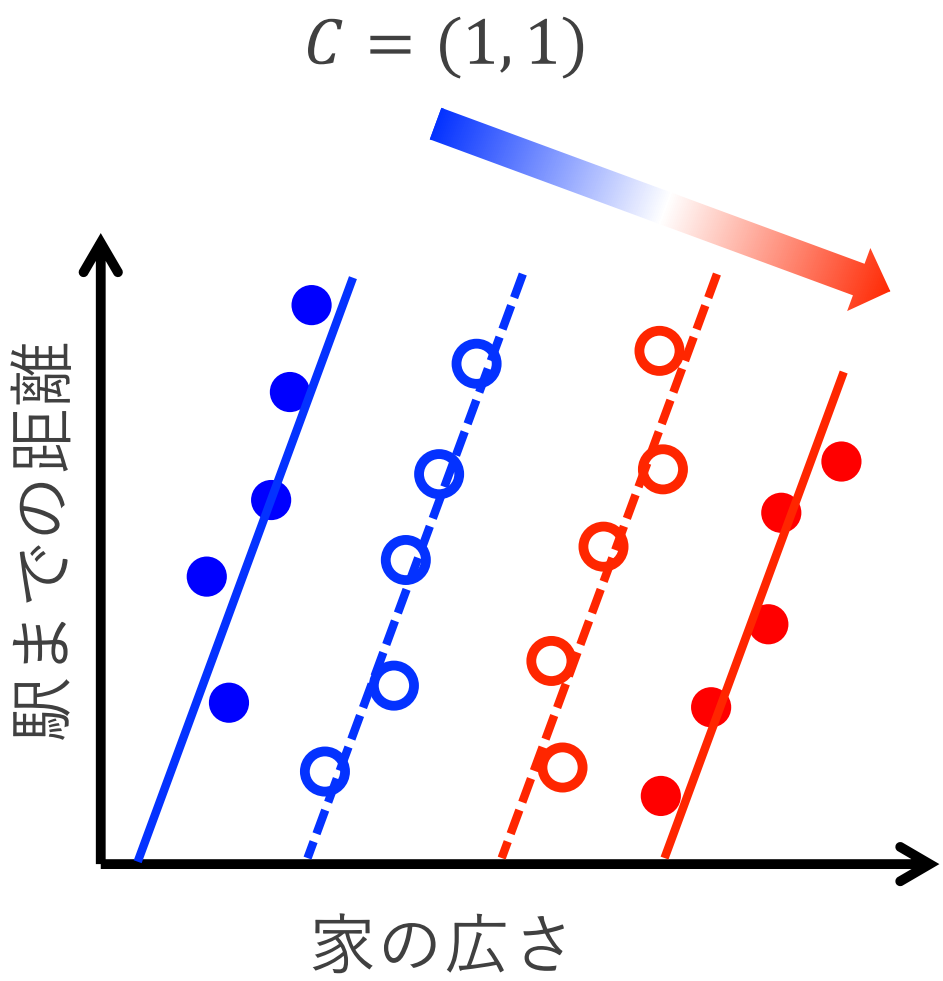
→ 未知データを上手く予測すること（汎化性能）

$$y \approx w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

汎化性能を高めるためには，必要な特徴量を見極めることが重要となる

特徴量選択の効果①



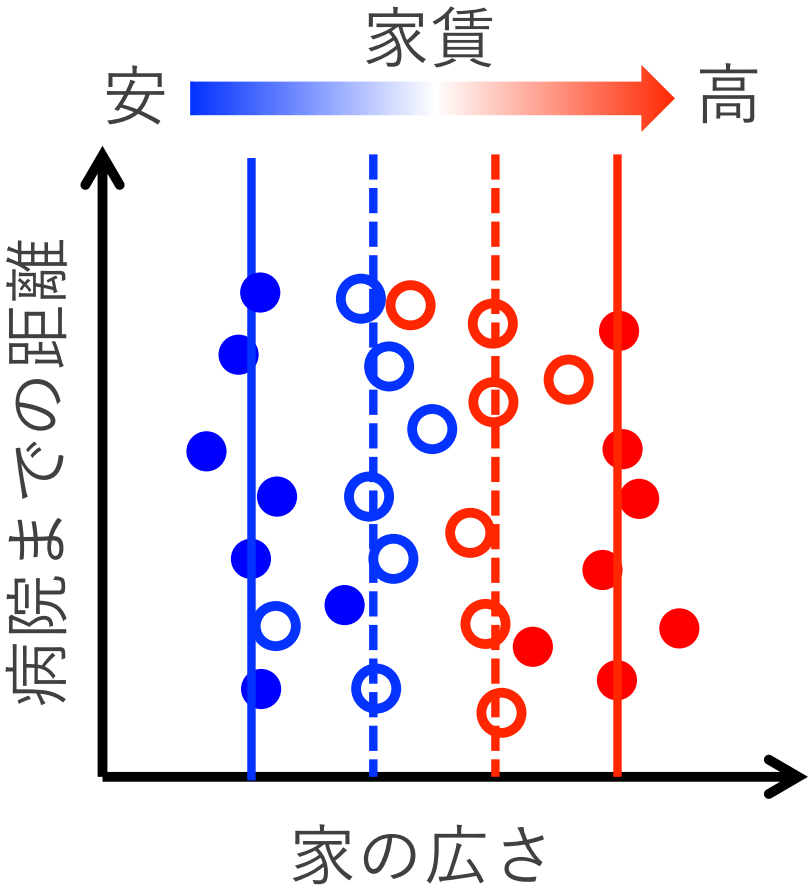


$Error(1,0) > Error(1,1)$

- 両方とも予測に必要

特徴量選択の効果②

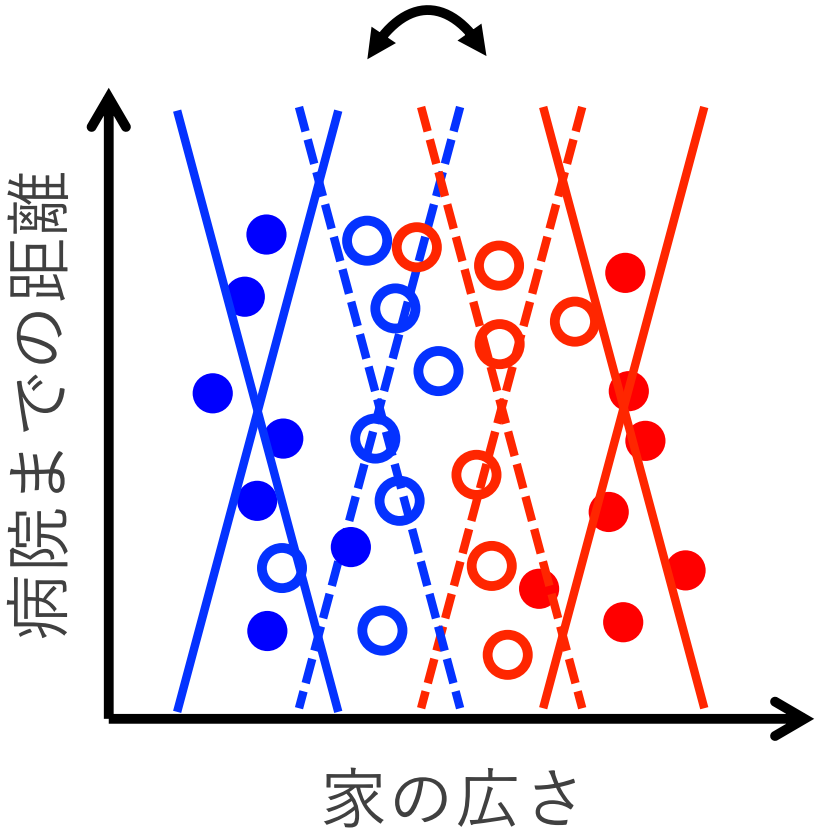
家の広さのみ $C = (1, 0)$



$Error(1, 0) < Error(1, 1)$

家の広さと病院までの距離

$C = (1, 1)$



- 家の広さ　　：必要
- 病院までの距離：不要

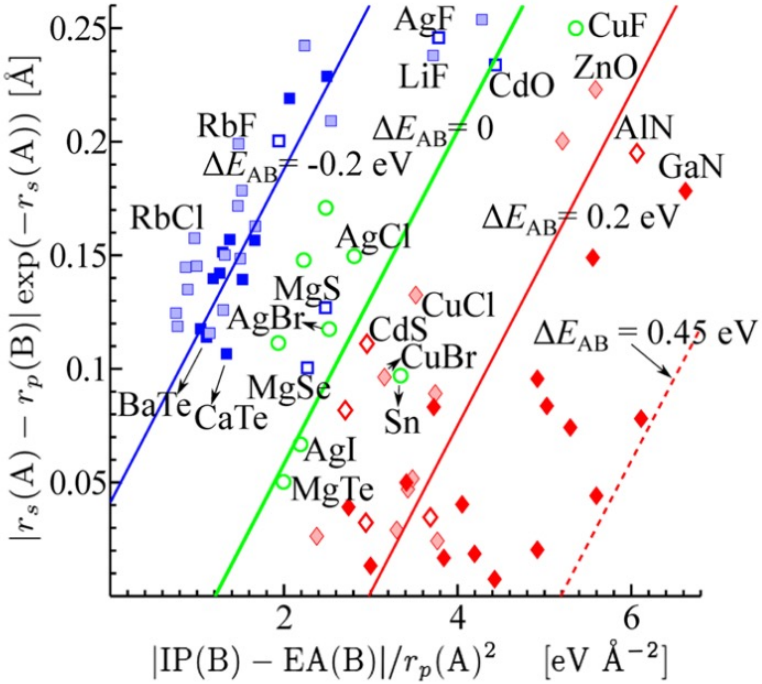
物理学における特徴量選択の重要性

- 物理的知見の抽出
  - 選ばれた特徴量と目的変数の関係性を考察することで，物理現象への知見が得られる
    - Ghiringhelli *et al.*, 2015, Igarashi *et al.*, 2018
- 特徴量の準備の省力化
  - 選ばれた特徴量のみ準備すればよいため，実験の計測時間や回数及び数値シミュレーションの回数を削減することが可能
    - Nakanishi-Ohno *et al.*, 2016

材料科学における特徴量選択の応用

1. 電池材料に関する化学反応の安定性予測
   Sodeyama *et al*, 2018

2. 半導体化合物の結晶構造予測
   Ghiringhelli *et al*., 2015(右図)



2つの特徴量からなる線形モデルによる結晶構造予測

内容

- 機械学習と特徴量選択
- 全状態探索法とWeight Diagram
- 特徴量選択の信頼度評価
- 拡張された特徴量空間における特徴量選択

データの表記

- 目的変数: $y \in \mathbb{R}$
- 特徴量: $\boldsymbol{x} = \left(x_1, x_2, \ldots, x_p\right)^T \in \mathbb{R}^p$

N 個の観測データ

$$\boldsymbol{y} = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$$
$$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T \in \mathbb{R}^{N \times p}$$

（補足）目的変数 y と各特徴量 $x^{(i)}$ は正規化等の前処理済みとする

最小二乗法による線形回帰

真のモデルを線形関数でモデル化：

$$f(\boldsymbol{x}; w_0, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_p x_p$$
$$= w_0 + \boldsymbol{w}^T \boldsymbol{x}$$

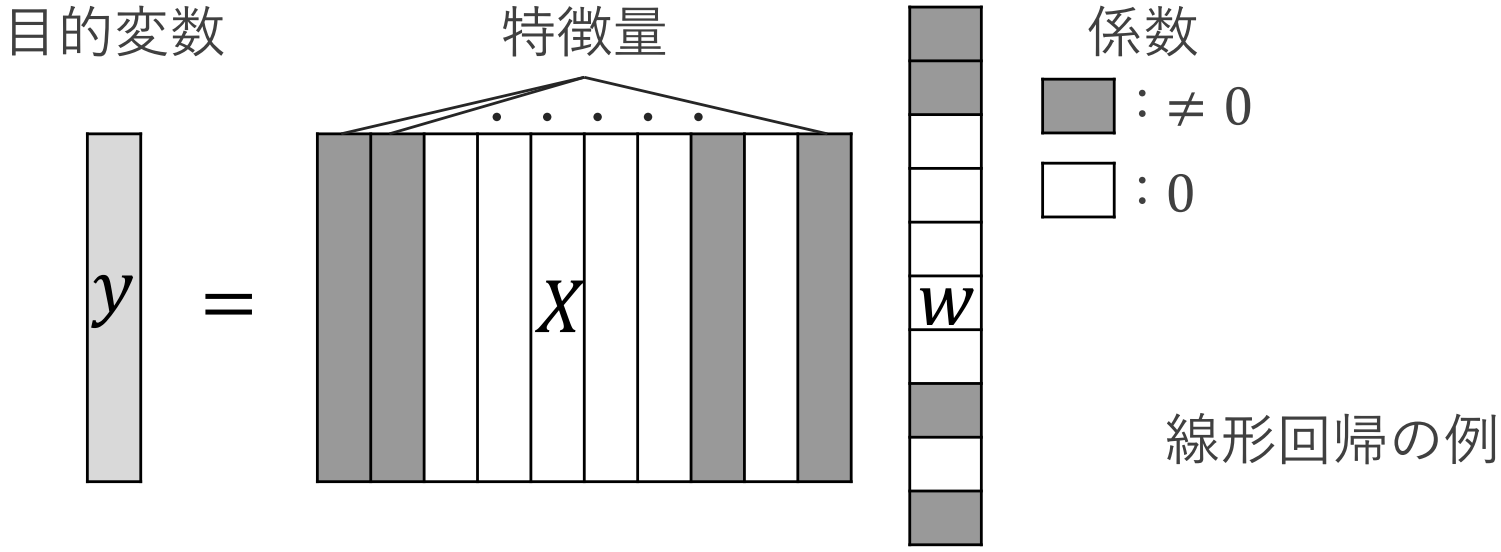
- 切片：$w_0 \in \mathbb{R}$
- 係数：$\boldsymbol{w} \in \mathbb{R}^p$

平均二乗誤差の最小化により，係数及び切片の値を求める

$$E(w_0, \boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} \left(y_n - f(\boldsymbol{x_n}; w_0, \boldsymbol{w}) \right)^2$$

特徴量選択

目的変数を説明するために重要な特徴量のみを選び出すことを目的とした，統計的機械学習の一分野



線形回帰の例

代表的な手法

- Lasso (Tibshirani, 1996)
- **全状態探索** (Exhaustive search) (Igarashi et al., 2018)

インジケータベクトルによるモデルの指定

インジケータベクトル $\boldsymbol{c}$ によって部分モデルを定義

$$\boldsymbol{c} = (1, 1, 0, \ldots, 0, 1, 0, 1) \in \{0,1\}^p$$

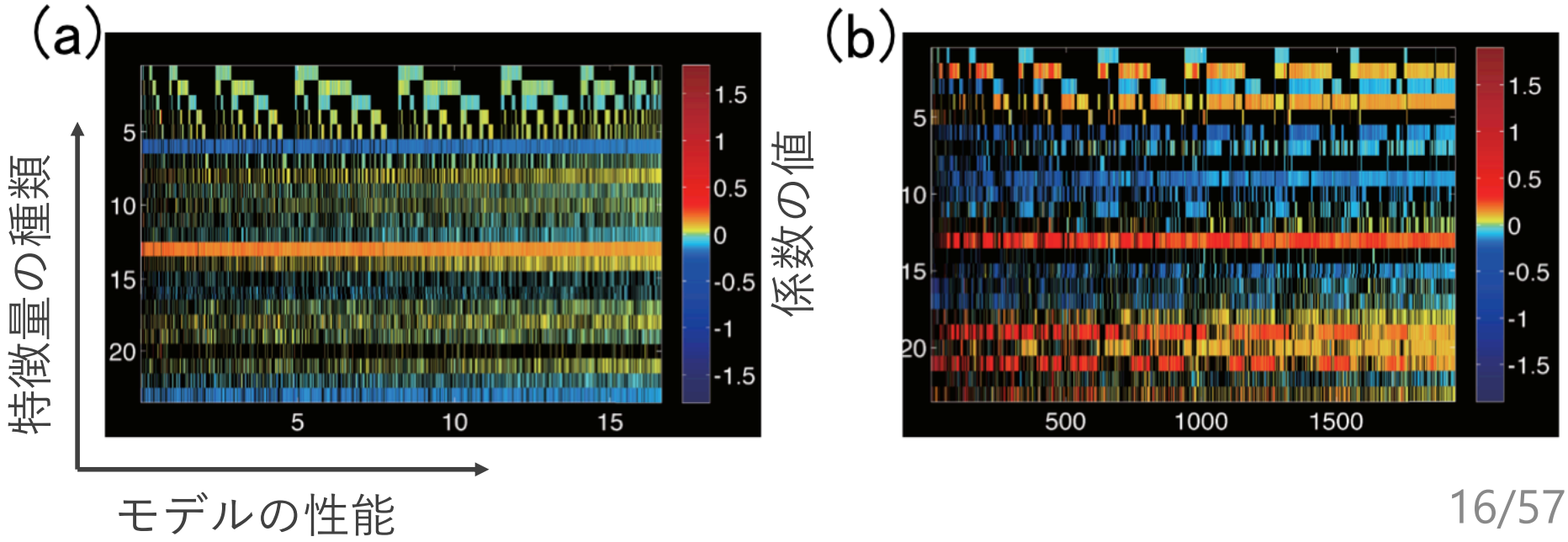
- i 番目の要素が特徴量 i に対応
  - 1:モデルに含まれる
  - 0:モデルに含まれない
- インジケータベクトルは 2^p 状態を取る
  - その内1つは空のモデル

厳密な特徴量選択を行うにはどの手法であっても計算量が指数関数的に増加する(Cover and Van Campenhout, 1977)

全状態探索とWeight Diagram

全状態探索：c の組み合わせを全て調べ，性能の良いモデルを探索する枠組み

探索結果から性能上位のモデルの係数をプロットする (Weight diagram) ことで，特徴量選択の揺らぎを定性的に分析する (Y. Igarashi *et al.* 2018 J. Phys.)



p（変数の数）が20程度まであれば厳密な全状態探索が適切 (1/3)

市川寛子
東京理科大

桑谷 立
JAMSTEC

五十嵐康彦
筑波大

p（変数の数）が20程度まであれば厳密な全状態探索が適切 (2/3)

- 脳科学/行動科学
  - NIRS: ASDとADHDの鑑別
    - Ichikawa *et al.*, 2014 Front Hum Neurosci
  - 歩行動作からASD症状の強さの回帰
    - Shigeta et al., 2018 Adv Biomed Eng
- 地球科学
  - 高精度な津波堆積物の判別を実現
    - Kuwatani *et al.*, 2014 Scientific Reports
  - 含水マグマの熱力学モデルを構築
    - Ueki *et al.*, 2020 Physics of the Earth and Planetary Interiors,
  - 岩石流体相互作用における反応経路を全探索によって決定
    - Oyanagi *et al.*, 2021 The European Physical Journal
  - その他　3本

p（変数の数）が20程度まであれば厳密な全状態探索が適切 (3/3)

- 物質材料科学
  - リチウムイオン電池の電解液材料探索への応用
    Sodeyama *et al.*, 2018 Phys. Chem. Chem. Phys, その他3本
  - 高収率なナノシート合成開発への応用
    Nakada *et al.*, 2019 Adv. Theory & Sim. その他4本
  - リチウムイオン二次電池用 有機負極活物質の探索
    Nakada *et al.*, 2019 Adv. Theory & Sim. その他3本
  - 熱応答性高分子の機能予測モデルの開発
    Hiruta *et al.*, under review
- 核融合
  - トカマク型核融合炉JT-60Uにおける高ベータ障害予測モデル構築
    Yokoyama *et al.*, 2019 Fusion Engineering and Design その他1本
  - 大型ヘリカル装置における放射性崩壊回避のためのデータ駆動型制御
    Yokoyama *et al.*, 2022 Plasma and Fusion Research